Interactive Visualization Systems and Data Integration Methods for Supporting Discovery in Collections of Scientific Information

A Thesis

Submitted to the Faculty

of

Drexel University

by

Donald Anthony Pellegrino Jr.

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

May 2011

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Interactive Visualization Systems and Data Integration Methods for Supporting Discovery in Collections of Scientific Information	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Drexel University,Philadelphia,PA,19104**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Technological developments have been enabling additional sharing and reuse of scientific information. Current indexing methods support query-based search and filtering, however they do not support overviews and exploration. Due to these limitations of existing indexing methods, it is challenging to discover records and connections that relate information in new and potentially insightful ways. We developed prototype systems and computational methods for integrating collections from multiple sources within a domain into a single, unified graph data structure. Graph-theoretic measures and visualizations were then applied to identify relations and records that support discovery tasks. Three collections of molecular information were studied: (1) influenza protein sequences from the National Center for Biotechnology Information, (2) Open Notebook Science notebooks and databases from Drexel University and other academic chemical research laboratories, and (3) project data from drug discovery projects at Pfizer R&D. We designed methods for data integration within these collections. We then analyzed the integrated collections to design interactive visual tools and computational methods that could systematically identify relations and records that have a high potential to lead to novel discoveries in these areas. We conducted interviews with domain experts to evaluate the effectiveness of these designs. These studies demonstrate the feasibility of the new indexing methods to improve the discoverability of novel connections across multiple collections within a domain.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**118**	

Dedications

This dissertation is dedicated to my wife Jill and to my daughter Madelyn for their love, support, and patience.

Acknowledgments

My studies in the Information Studies PhD program were supported by many organizations. I was supported by the College of Information Science and Technology at Drexel University. Work was also supported by the National Visualization and Analytics Center, a U.S. Department of Homeland Security program operated by the Pacific Northwest National Laboratory (PNNL). PNNL is a U.S. Department of Energy Office of Science laboratory. This research was supported in part by the National Institutes of Health through resources provided by the National Resource for Biomedical Supercomputing (P41 RR06009), which is part of the Pittsburgh Supercomputing Center. This research was supported in part by the National Science Foundation through TeraGrid resources provided by the National Center for Supercomputing Applications at the University of Illinois under grant number TG-IRI100001. I was supported in part by the ipl2 project at the College of Information Science and Technology at Drexel University. I was supported in part by the ACIN Warfighter Support Program of the Applied Communications and Information Network (ACIN) Center though a partnership between the U.S. Army Communication-Electronics Research, Development and Engineering Center (CERDEC) and Drexel University. I was also supported in part by Pfizer Inc. through the Pfizer-Drexel Collaboration.

I would like to thank the members of the DHS Student and Alumni Network as well as the participants and organizers of the Visual Analytics Summer Camp at PNNL. I am particularly grateful for the time Jim Thomas and Kristin Cook took with me to introduce me to Visual Analytics research and the Visual Analytics community.

During the course of my studies, I had the opportunity to work with Christopher Cannon, and William Regli at the ACIN Center and Giovanni Oddo at CERDEC. They have provided excellent

guidance on carrying out a successful research project. I am very grateful to Tom Hewett for sharing his expertise and for teaching me how to approach and carry out the process of research for a sponsor.

I would like to thank the members of the North-East Visualization and Analytics Center at the Pennsylvania State University, including Chi-Chun Pan, Prasenjit Mitra, Anthony Robinson, Michael Stryker, Ian Turton, and Junyan Luo. I am particularly grateful for the time and support that both Alan MacEachren and Chris Weaver provided to contribute to the team and to guide me individually.

I am grateful to the members of Drexel University's Center for Integrated Bioinformatics, particularly Aydin Tozeren and Will Dampier. Thanks also go to William Doran for his assistance with the compute cluster in Drexel's College of Information Science and Technology.

Our collaboration with Pfizer R&D has been an excellent opportunity. I am grateful to the Pfizer researchers who have made this collaboration a success, including Alan Mathiowetz, David Anderson, and Rishi Gupta. Bruce Lefker has been instrumental in making the collaboration possible and providing guidance. Jared Milbank's technical skill and insightful approaches have been particularly valuable. I have learned a lot from him and the team.

I would like to thank Jean-Claude Bradley for inviting me to work with him and his laboratory and for introducing me to the open-science community. His student Evan Curtin was very gracious in allowing me to observe activities in the laboratory and explaining processes to me. Andrew Lang from the ONS community contributed a very useful experiment to the Open Notebook Science Challenge site for visualizing ONSC data.

Without the support of my committee, this dissertation would not have been possible. Robert Allen fed me with new ideas starting with his course “Topics in Information Retrieval, Visualization, and Bibliometrics: Multimedia Information Retrieval” and throughout the duration of my studies. He was always available with an open door for discussions. Xia Lin guided me on my work with the ipl2 and provided some of the foundational contributions to the field of Bibliometric Visualization. Longjian Liu encouraged me to think constantly about the impact to the research scientist and the benefit of these efforts. Jean-Claude Bradley was supportive with both his time and ideas. During our weekly meetings, he helped me keep in touch with the developments in Open Notebook Science. This work would not have been possible without taking a multi-disciplinary approach. It would not have been possible for me understand Open Notebook Science as Chemists see it or to really understand it without his deep support.

Finally, I must acknowledge my Committee Chair and PhD Advisor Chaomei Chen. Chaomei Chen has been my mentor throughout my studies and I am grateful to have worked with such a great master of the profession. As an MSIS student, I had the opportunity to perform an independent study under his guidance and the experience inspired me to pursue a PhD. Through his mentorship, I have learned to read the scientific literature and even to begin to contribute to it. A great many of the opportunities that I have had were a result of his leadership.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Interactive Visualization Systems and Data Integration Methods for Supporting Discovery in Collections of Scientific Information

Donald Anthony Pellegrino Jr.

Chaomei Chen, Ph.D.

Technological developments have been enabling additional sharing and reuse of scientific information. Current indexing methods support query-based search and filtering, however they do not support overviews and exploration. Due to these limitations of existing indexing methods, it is challenging to discover records and connections that relate information in new and potentially insightful ways. We developed prototype systems and computational methods for integrating collections from multiple sources within a domain into a single, unified graph data structure. Graph-theoretic measures and visualizations were then applied to identify relations and records that support discovery tasks. Three collections of molecular information were studied: (1) influenza protein sequences from the National Center for Biotechnology Information, (2) Open Notebook Science notebooks and databases from Drexel University and other academic chemical research laboratories, and (3) project data from drug discovery projects at Pfizer R&D. We designed methods for data integration within these collections. We then analyzed the integrated collections to design interactive visual tools and computational methods that could systematically identify relations and records that have a high potential to lead to novel discoveries in these areas. We conducted interviews with domain experts to evaluate the effectiveness of these designs. These studies demonstrate the feasibility of the new indexing methods to improve the discoverability of novel connections across multiple collections within a domain.

CHAPTER 1: INTRODUCTION

Summary

We use three studies to explore the method of integrating data into a semantic network, visualizing the network, and highlighting nodes with structurally interesting characteristics. These methods demonstrate feasible approaches to indexing that can address the volume and diversity of data being produced by recent technological advancements in data production. The heterogeneous semantic network has the characteristic of supporting interactive visual projection for exploratory analysis. It also enables quantitative analysis. We show that this combination can be used to assisting users with the identification of key records and relationships.

Databases of structured data make use of formal data types and often use storage that is optimized for the contained types. Examples of data types include enumerations, integers, decimal values, strings, and Booleans. Here the term database is used to refer to repositories for structured data, as opposed to databases containing text, documents, articles and other literature. Literature is used to refer to data that has the special data type of text. Collections of literature used for linguistic or statistical content analyses are commonly referred to as corpora. The single collection of literature used may be referred to as a corpus. We can use this terminology to describe a hierarchy with databases of data on one side, corpora of literature on the other, and collections as a broader term for elements of content from databases, corpora, or aggregated sets from both.

Although literature is often stored in a Relational Database Management System (RDBMS) using a formal structure, it represents a special case of data. Data are generally easily decomposed into very granular instances. In many textbooks, information is considered as a higher-level

concept than data. Text decomposes in a different way, such that letters, words, sentences, paragraphs, sections, and articles have different conceptual semantics than tuples composed of integers, floats, strings, enumerations, and other traditional data types. Databases relate data in fundamentally different ways than corpora of literature relate articles and their contents. Tuples are often related to each other in a way that is conceptually different from the way that articles are related to each other in a corpus. Relating tuples to articles and thereby connecting databases and collections of literature is an open problem. These methods described here show how tuples can be systematically related to articles in the context of specific domains and collections.

Due to the differences in how collections of literature and databases of data are searched and explored it is challenging to make novel connections that relate information stores as data with information contained in the literature. One specific mechanism of scientific discovery is the creative process of making novel connections between previously disconnected bodies of knowledge (Swanson 1986; Fleming, Mingo et al. 2007). While there may be many ways to relate data and literature, the objective of these new methods is to relate them in such a way that they can provide systematic support for the creative process of making novel connections.

Presentation

In “CHAPTER 2: Literature Review,” publications in information science and other fields are summarized and compared. Two histories are traced. First, in “The Scientific Research Information Environment,” the nature and kind of artifacts produced by the scientific research process are examined. The impact of the Internet on scientific communications artifacts is explored by comparing classification schemes produced in 1971 and 2003. The 2003 scheme is then interpreted in terms of recent developments including, Open Notebook Science, Digital

Library Systems, and Cyberinfrastructure. The problem of modern Information Overload is considered along with the way it has historically been addressed – by revisiting indexing strategies.

After first examining the historical and modern Scientific Research Information Environment, we then look at techniques that have been used to work within that environment. These are described in “Navigating the Environment.” Comparisons are made between the seminal indexing work of Eugene Garfield made possible by the ISI citation index and the indexing work of Lawrence Page made possible by the hyperlink structure of the World Wide Web. We then examine discovery algorithms in “Literature Related Discovery (LRD).” The term “Data Related Discovery” is introduced to contrast with LRD and to examine new developments in data mining and heterogeneous data analysis. The role of visualization in analysis of data is described along with a brief account of the new field of Visual Analytics. Cognitive aspects of human creativity and its relationship to visualization are described in “Scenario Visualization.”

In “CHAPTER 3: Study Design,” we transition from the identification of current needs to approaches for addressing those needs. With recognition of a need to integrate collections of literature and collections of data, we then explore methods to elaborate the nature of the engineering problems and the human factors involved. In this chapter, we define the specific aims of the studies, the research questions, and the methodology. We also examine the lessons learned from our involvement with the 2008 VAST Challenge. This chapter outlines and describes the high-level design and the goals of the three studies that follow.

In “CHAPTER 4: Influenza Protein Sequence Analysis,” we describe methods and a prototype system for mapping all of the available Influenza protein sequence data published by the National Center for Biotechnology Information. This study investigates the engineering problems

of constructing a large graph from a real-world scientific data set. We explore the domain-specific issues of NCBI protein sequence records. We also explore the utility provided by a full visualization of the data and compare this to the capabilities of the user interfaces provided by multiple published systems. Animation is used to explore macroscopic patterns of the swine flu pandemic.

“CHAPTER 5: Open Notebook Science” reports on studies done in the domain of Open Notebook Science. We examine the nature of the literature and data collections within this domain. We look at ways in which they can be combined. Visual representations of the data are discussed along with reactions to those representations by researchers.

“CHAPTER 6: Pfizer Drug Discovery Projects” reports on studies performed within the domain of drug discovery. Project data from drug discovery projects run by Pfizer Research and Development are analyzed. The data collection is composed of compound structure similarity for compounds synthesized during the course of a drug discovery project. The literature collection employs PowerPoint slides produced by the project team. We created graphs and their projections to represent the collection of project data and information extracted from the slides. Graph-theoretic measures were used to identify compounds and links of interest. An interview was conducted with a Pfizer researcher to evaluate the relevance of the representations and measures for providing insight into the project.

In “CHAPTER 7: Conclusions and Discussion,” we summarize the conclusions that can be drawn from the studies. We also provide a discussion of possible future directions and opportunities for further study. The three studies of Influenza protein sequence data, Open Notebook Science, and drug discovery demonstrate specific methods for exploring data in new ways that support discovery.

CHAPTER 2: LITERATURE REVIEW

Observing the current trends in cyberscholarship, digital library systems, and Open Notebook Science, we can see the shape of a future scientific environment emerging. Features of this environment include the inclusion of research data alongside research publications and the availability of artifacts from all points along the scientific process. Pieces of this environment are being built opportunistically by researchers who are taking advantage of the tools on-hand while other pieces are being deliberately engineered with support from the National Science Foundation (NSF), the National Academies Board on Research Data and Information (BRDI)[1], and other large funding agencies. Tools for exploring scientific literature have been developed in the traditions of bibliometrics and literature related discovery. Separately, analytical tools are being developed to address issues of information overload. Opportunities exist to combine techniques from these efforts to support development of tools for navigating within the broader scientific research information environment.

The Scientific Research Information Environment

Cyberscholarship, Cyberinfrastructure, Discovery and Innovation

Cyberscholarship refers to "new forms of research and scholarship that are qualitatively different from tradition ways of using academic publications and research data (Arms and Larsen 2007)." Often, issues in cyberscholarship are associated with issues in cyberinfrastructure. Infrastructure refers to software tools and hardware platforms that facilitate these new forms of research and scholarship. Discussions of the cyberinfrastructure often include free and open-source software systems running on supercomputers that are

[1] Homepage for the Board on Research Data and Information at the National Academies: http://sites.nationalacademies.org/pga/brdi/index.htm.

connected via the NSF TeraGrid.[2] NSF also has a program for "Cyber-Enabled Discovery and Innovation:"

> *"Cyber-Enabled Discovery and Innovation (CDI) is NSF's bold five-year initiative to create revolutionary science and engineering research outcomes made possible by innovations and advances in computational thinking. Computational thinking is defined comprehensively to encompass computational concepts, methods, models, algorithms, and tools. Applied in challenging science and engineering research and education contexts, computational thinking promises a profound impact on the Nation's ability to generate and apply new knowledge. Collectively, CDI research outcomes are expected to produce paradigm shifts in our understanding of a wide range of science and engineering phenomena and social-technical innovations that create new wealth and enhance the national quality of life (Misawa, Russell et al. 2009)."*

A vision for the National Cyberinfrastructure was developed collaboratively and is articulated in (National Science Foundation Cyberinfrastructure Council 2007). Execution of this vision is being managed by the Nation Science Found Office of Cyberinfrastructure.[3] The work described here complements these initiatives and could ultimately contribute to the National Cyberinfrastructure in the form of search algorithms and data exploration techniques. The NSF also established an Advisory Committee for Cyberinfrastructure (ACCI) composed of six task forces. The task forces are:

[2] http://www.teragrid.com
[3] http://www.nsf.gov/dir/index.jsp?org=OCI

- Campus Bridging
- Cyberlearning and Workforce Development
- Data and Visualization
- Grand Challenges
- High Performance Computing
- Software for Science and Engineering

These task forces collected feedback from academia and industry to identify needs and describe visions of a future cyberinfrastructure. The final reports from these task forces were published on April 1, 2011,[4] including the Final Report from the Task Force on Data and Visualization (Atkins, Baker et al. 2011).

General Digital Library Systems

The library communities are seeing a trend in the shift from the dominance of physical local collections to digital federated collections of resources (Smith 2009). Professional curation of digital collections requires much more technology and process than is directly supported by the popular file system tools available today. Extensive support for metadata is one discriminating characteristic separating digital library systems from simple file systems. Digital library systems provide a technology layer to support professional collections management on top of hierarchical file system and relational database management systems technologies. Two general-purpose digital library systems are the Flexible Extensible Digital Object Repository Architecture[5] (FEDORA) and DSpace.[6] In May of 2009 the organizations that supported the development of FEDORA and DSpace merged to unify their efforts as the DuraSpace

[4] http://www.nsf.gov/od/oci/taskforces/
[5] FEDORA Homepage: http://www.fedora-commons.org
[6] DSpace Homepage: http://www.dspace.org

Organization (Morris, Kimpton et al. 2009). The ipl2,[7] a combination of the Internet Public Library (IPL) and the Librarians' Internet Index (LII) currently uses the FEDORA system to maintain records.

By taking a systems approach, digital libraries generally include a server component for managing a collection. Container file formats can be considered a lightweight approach to metadata management. By utilizing a server process digital library systems impost a form of centralized control that is implicitly assigned to the organization running and managing the server process. Container file formats such as MPEG4[8] and HDF5[9] manage data by using wrapper architectures. In general, the data to be managed is encapsulated in a layer of metadata that describes the data. Multiple layers are enclosed in a single file. While the FEDORA and DSpace models also use multiple layers, a differentiating characteristic of container file formats is that the metadata travels with the data rather than existing on a server as a reference. This allows for a decentralized approach to metadata management. Issues arising from incompatible data formats may consume ninety percent or more of the time spent on a data visualization project and increased usage of HDF5 or the newly proposed F5 container format are expected to help improve the situation (Benger 2009).

The familiar examples of Microsoft Windows Media Player (WMP) and Apple iTunes illustrate the difference between the centralized approach of digital library systems and the distributed approach of container file formats. Windows Media Player provides both search and browse capabilities. WMP makes use of an internal index of the metadata for each song. This is the index used to construct trees for browsing or to find songs based on a search string. The user

[7] Ipl2 About: http://www.ipl.org/div/about/
[8] http://www.chiariglione.org/mpeg/standards/mpeg-4/mpeg-4.htm
[9] http://www.hdfgroup.org/HDF5/

cannot directly maintain the records in this index. Instead, metadata is stored in the file for each song. Each file type exposes its own metadata fields. For example, the ID3 metadata format is often used with the MP3 file format for music. Microsoft's Advanced Systems Format (ASF) is another type of media format that combines the audio streams for songs with a structured format for metadata into a single file. Windows Media Player reads the metadata from each song file in a user's media library to construct its index. To edit the metadata, a user must make changes that are compatible with the specific metadata format used by the media type for a given song file. When a media file is moved from one computer to another, the associated metadata travels with it. A user does not need to reenter the metadata when the file is moved into a new media library. A disadvantage of this approach is that some metadata fields may not be available for songs that are using more limited file formats. This demonstrates the utility of storing metadata in a container file format. Apple's iTunes uses a centralized metadata approach and allows users to maintain song metadata in a single master XML file. This has the advantage that it allows for consistency of metadata structure for all song records. However, when a song is moved from one iTunes library to another the metadata must be copied separately or reentered in the new library. This demonstrates the utility of storing metadata in a centralized location.

Domain Specific Digital Library Systems

While the DuraSpace projects seek to be as general as possible, domain specific digital libraries have also emerged. "The Software Environment for the Advancement of Scholarly Research (SEASR), funded by the Andrew W. Mellon Foundation, provides a research and development environment capable of powering leading-edge digital humanities initiatives (SEASR 2009)." SEASR, through its partnership with the National Center for Supercomputing Applications at the

University of Illinois at Urbana-Champaign, creates a bridge between humanities scholars and the supercomputing community. Supercomputing has historically focused on the needs of scientists working in the physical sciences. Despite this heritage, management of scientific data remains an unsolved problem even in the physical sciences. The Science Commons organization seeks to make scientific research data more reusable and more useful. "If we can *systematically* increase our chances of making big discoveries and decrease the likelihood that we are ignoring information that we should be using then that's the best chance we have to get these breakthroughs in understanding about our bodies and about drugs (Dylan 2009, emphasis added)."

Digital library systems come from a tradition of librarianship. Thus, even though a generalized object model is used, the core object model tends to be stylized off a card catalog, with records for books influencing the model. For example, FEDORA objects tend towards a metadata approach that includes at least authorship and title. This makes it difficult to fit collections of data records into the model, as tuples of data are not normally named and authorship may not have such priority. Alternative architectures are used for large collections of scientific data. Large, heterogeneous, collections of scientific data are today exemplified by the Global Earth Observation System of Systems (GEOSS),[10] and the Large Hadron Collider (LHC) data systems.[11]

Open Notebook Science

"Open Notebook Science is the practice of making the entire primary record of a research project publicly available online as it is recorded (Wikipedia contributors 2009)." Open Notebook Science shares the goal of publicizing research data that is advocated by the Science Commons. Open Notebook Science goes further and extends the objectives in the vein of the

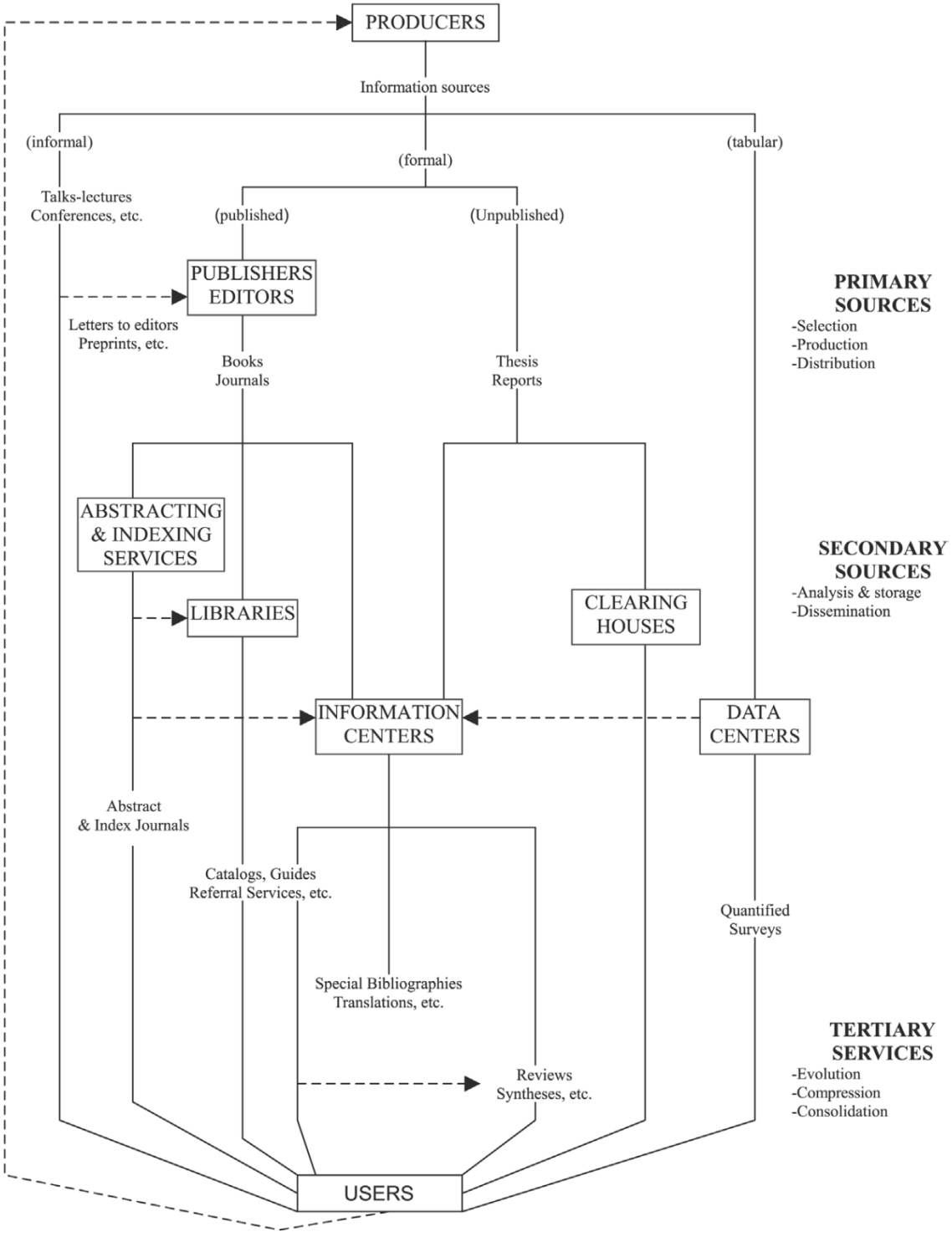
[10] http://www.earthobservations.org/geoss.shtml
[11] http://lhc.web.cern.ch/lhc/

open-source software development model to include both data and process. Exemplars of this approach include the "UsefulChem[12]" site by Jean-Claude Bradley and "The Synaptic Leap[13]" site by Matthew Todd. In many scientific institutions, laboratory notebooks are kept as paper records by researchers. These notebooks are often archived in institutional libraries. A primary use of these records has been to support the patent process. While institutions have attempted to digitize notebooks, opening them to the public as they are being written represents a significant change in process. Academic use of Open Notebook Science is described further in (Bradley, Lang et al. 2011).

Models of Scientific Communication

In 1971, the United Nations Educational, Scientific, and Cultural Organization (UNESCO) and the International Council of Scientific Unions (ICSU) cooperated to publish the UNISIST model of scientific and technical communications shown in Figure 1.

[12] http://usefulchem.wikispaces.com
[13] http://www.thesynapticleap.org



Note: Reproduced by permission of UNESCO

Source: UNISIST (1971, p. 26)

Figure 1: The original UNISIST model as reproduced in (Søndergaard, Andersen et al. 2003).

In 2003, revisions and updates to this model were proposed that extend it to include Internet-based scholarly information (Søndergaard, Andersen et al. 2003). The extension of the 1971 model in 2003 reveals that even the most general model of scientific communication in 1971 was insufficient to classify the communications that emerged with the introduction of Internet technologies. At the highest level, the model identified three types of communications; primary sources, secondary sources, and tertiary services as defined in Table 1.

Primary Sources	"Primary literature is the researcher's and knowledge producer's primary medium for claiming original findings, theoretical analysis, empirical data etc.: Monographs. . . Journal articles. . . Critical-analysing reviews. Conference presentations. 'Grey' literature. . .Patents. Standards. [p.318]" "Source literature is either literature produced in order to supply researchers with information (e.g. translation journals) or information produced to other purposes than research, but used as information by researchers (e.g. music and fiction)… *Data archives, Statistical documents, tabular documents* [p.319, emphasis added]."
Secondary Sources	"Secondary literature / bibliographical literature. This is literature that registers, describes and organises the primary literature as well as the other categories (including the secondary literature itself). Secondary information systems are the core focus of the library, documentation, and information science profession. Bibliography is a discipline that studies this area: Subject bibliographies and bibliographical databases. … [p.319]."
Tertiary Services	"Tertiary literature / review literature / 'outlines.' This is literature summarising and synthesising knowledge in the primary literature: Handbooks. . . Review

	articles. . .Data hand-books, tabular documents 2 (synthesising original statistical sources) [p.319-320]."

Table 1: High-level classifications in the UNISIST model.

These three types of communication were superimposed on a flow model that described the artifacts during transitions and stages of information moving from producers to users. The extended model placed the Internet alongside the entirety of the flow covering the full process from producer to user. It also added “Preprint Databases,” “Scientific and Research Organizations Servers,” and “Search Engines” as some of the significant new objects in the extended model (Søndergaard, Andersen et al. 2003, Figure 5, p.303). A final extension was to enclose the entire model within the boundary of a domain. This was done in recognition that different epistemologies in a given domain will emphasize different knowledge sources (Søndergaard, Andersen et al. 2003, p.305). The revised model is shown in Figure 2.

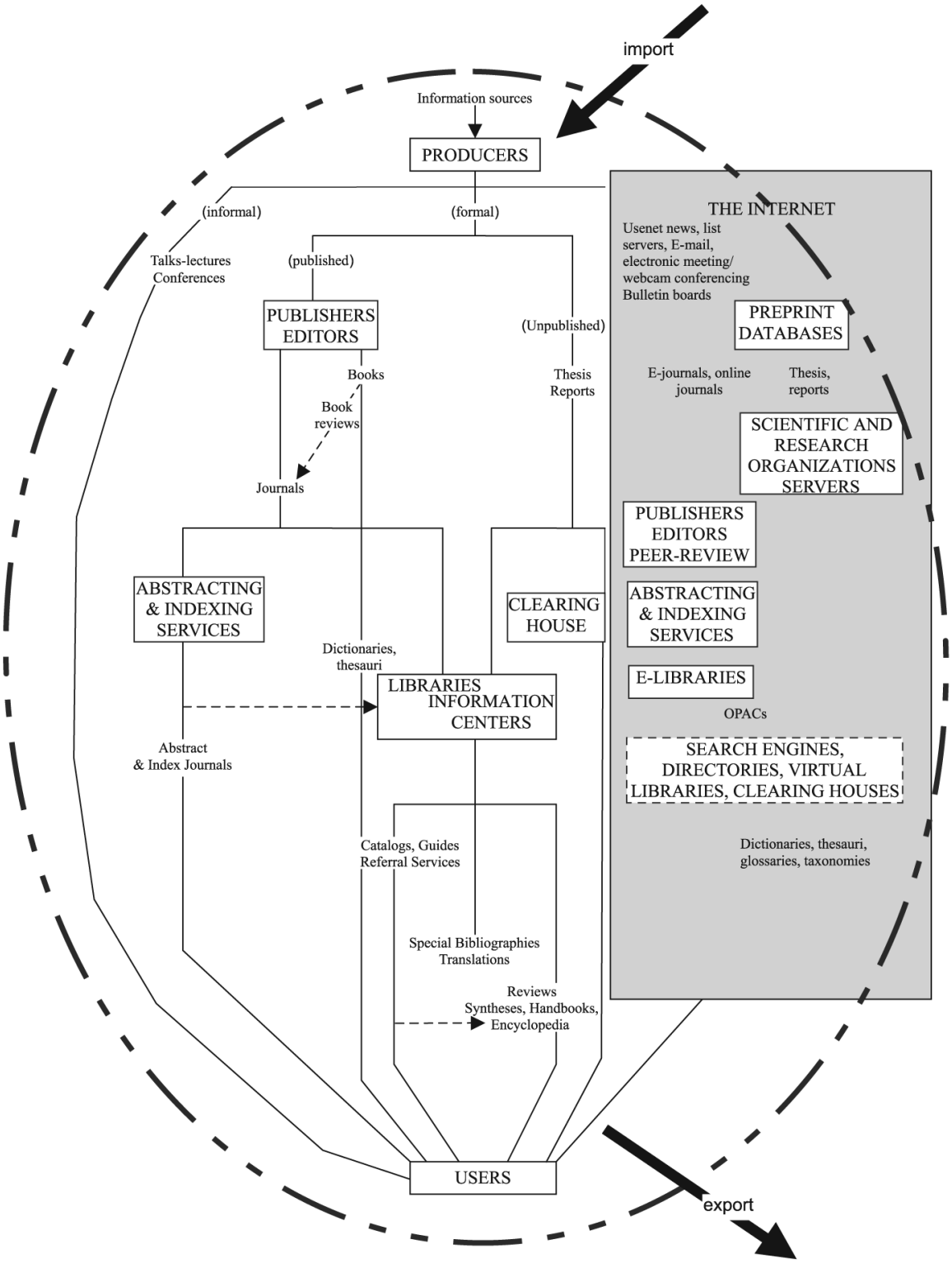


Figure 2: UNISIST model updated to reflect effects of the Internet (Søndergaard, Andersen et al. 2003).

Open Notebook Science as described above was not included in the 2003 revision, however it might have occupied a space in the upper-right of Figure 2, just above “Preprint Database.” The observation that different epistemologies in a given domain will emphasize different knowledge sources influences the design of the experiments performed for this work. We therefore studied domains of chemical knowledge. Subdomains were defined by coordinating the interests of researchers with the specific collections that dominated their information environment. Due to the emphases of different knowledge sources within subdomains, a generalized method for integration is likely to be less tractable and less impactful to the researcher than methods that account for differing emphases.

Information Overload

The problem of too much information, or information overload, has become well recognized in popular culture. The issue pervades even personal information management such as email, tweets, and Facebook updates (Zeldes 2009). IDC predicted, “… in 2011, the amount of digital information produced in the year should equal nearly 1,800 exabytes, or 10 times that produced in 2006. The compound annual growth rate between now [2008] and 2011 is expected to be almost 60% (Gantz, Chute et al. 2008).” While the sheer volume of modern digital information creation is impressive, the problem of managing large collections is perennial. One historical response to a sudden large increase in data volume has been to revisit indexing strategies as described in an account of managing intelligence data during World War II: “The indexes were not started as part of a great documentation plan, but simply emerged as response to the continuing and rapidly growing problems of controlling vast amounts of intelligence consequent on the successes in breaking Enigma and other encryption systems (Brunt 2005).”

Navigating the Environment

Bibliometrics

Traditionally, physical libraries have been primarily concerned with the management of their own local collections. This includes providing local cataloging and indexing services. With interlibrary loan programs, the scope of resources exposed to a patron are expanded to include the collections of collaborating libraries. Access to such multi-institutional collections ranges in simplicity from searching each library's catalog individually, searching each library's catalog with federated search, or use of a single catalog that contains aggregate data from all of the collections. Specialty indexes and manually authored domain-specific bibliographies provide a problem or domain view into the literature independent of aggregations by physical collection.

Eugene Garfield founded the Institute for Scientific Information (ISI) in 1960 which produced subject-specific indexes of the literature (Garfield 2009). These indexes were unique in including the references cited by the articles included in the index. As these indexes became available electronically, large-scale analysis of the citation patterns became feasible. This fed the field of bibliometrics, which led to the development of algorithms and visualization systems. Examples of modern systems include HistCite (Garfield, Pudovkin et al. 2003), CiteSpace (Chen 2006), and AuthorLink (Lin, White et al. 2003). Each of these systems provides a perspective on the literature that is algorithmically derived from the citation data, as opposed to manual definition by expert bibliographers and index authors. In many cases, the algorithmically identified perspectives invite users to discover novel relationships and new insights regarding the problem, domain, or author being studied.

PageRank

The World Wide Web introduced a new multi-institutional document collection. This collection however lacks the professional curation and indexing practices followed by librarians. Search engines have attempted to expose the collection to users by generating their own indexes of the content and then providing a custom interface to the index. In 1998 Lawrence Page, co-founder of Google, filed a patent for the PageRank algorithm (Page 2001). The "field of the invention" is documented as:

> *"This invention relates generally to techniques for analyzing linked databases. More particularly, it relates to methods for assigning ranks to nodes in a linked database, such as any database of documents containing citations, the world wide web or any other hypermedia database (Page 2001)."*

Page's contribution recognized that indexing the World Wide Web could be seen as an extension of citation analysis and bibliometrics by interpreting hyperlinks as bibliographic citations. Indeed Garfield's *Science* article (Garfield 1972) is cited as a reference in the PageRank patent.

Within the field of bibliometrics, the unit of analysis is limited to the bibliography of a work. More abstractly, however the field has dealt with the connectedness of people and their ideas by using the measure of a citation as an indicator for behavioral phenomena regarding social networks, the formation of ideas over time, and the current state of an intellectual domain. In a sense then the limits on bibliometric analysis are an artifact of the materials generally curated by traditional libraries and their subsequent indexing by ISI.

The trends identified in the Scientific Research Information Environment point to an expansion of the types of media collected and an increase in digital libraries. In particular, data and supplementary materials from research works are increasing in availability. The unique contribution of ISI in 1960 to index a measure of connectedness through references cited can be compared to the contribution of PageRank in 1998 to index the connectedness through hyperlinks of web pages. Each of these seminal contributions was heavily influenced by the nature of the media being indexed. In journal articles, the citation serves as a behavioral indicator of intellectual constructionism. In web pages, the hyperlink serves as an analogous behavioral indicator. The metadata components of digital library systems and container file formats provide the opportunity to build networks of connected artifacts that transcend the explicit linkages established by journal articles through citation and web pages through hyperlinking.

Literature Related Discovery

Don Swanson pioneered the field of Literature Related Discovery in 1986 with the publication of "Undiscovered Public Knowledge" which opened:

> *"Knowledge can be public, yet undiscovered, if independently created fragments are logically related but never retrieved, brought together, and interpreted. Information retrieval, although essential for assembling such fragments, is always problematic. The search process, like a scientific theory, can be criticized and improved, but can never be verified as capable of retrieving all information relevant to a problem or theory. This essential incompleteness of search and retrieval therefore makes possible, and*

plausible, the existence of undiscovered public knowledge. … (Swanson 1986).”

The essential incompleteness of search and retrieval referred to by Swanson was explored in detail in 1968 by the philosopher Patrick Wilson with “Two kinds of power: an essay on bibliographical control (Wilson 1968).” Although Wilson’s work is not cited in “Undiscovered Public Knowledge” Swanson succeeds in integrating Wilson’s ideas with Karl Popper’s critique of positivism from the 1934 “Logik der Forschung” (The Logic of Scientific Discovery).

Swanson used computational analysis of citation data to infer syllogistic relationships between clusters of medical literature. It is notable that he used Garfield’s ISI Science Citation Index via the DIALOG system for his work. With his approach, he “… demonstrated that, at least qualitatively, the most successful attempts to treat Raynaud’s syndrome tend to produce the same effects on certain blood parameters that dietary fish oil has been claimed to produce. … (Swanson 1986).” This analytically discovered connection was then used as an initial hypothesis to be experimentally validated. It was later shown that fish oil did indeed alleviate the symptoms of Raynaud’s syndrome. For his work in Literature Related Discovery, Swanson received the ASIS&T Award of Merit in 2000, the highest honor given by the American Society for Information Science and Technology (Swanson 2001). In his acceptance speech he remarked “Among all the people whose writing have influenced and inspired me, an astonishingly high proportion of them have received an ASIS&T award, among them … Eugene Garfield … (Swanson 2001).”

Work in Literature Related Discovery (LRD)has continued with Smalheiser (Swanson and Smalheiser 1999) and Kostoff (Kostoff 2009) making notable contributions. The Arrowsmith (Swanson 2008) system attempts to capture and expose much of the algorithmic work. The

majority of the work in the field has continued to focus on the medical domain and by definition; all of it continues to use the literature as the primary unit of analysis, although some of the algorithms take advantage of the Medical Subject Headings (MeSH)[14] and domain ontologies to support natural language processing (NLP) aspects. A comprehensive review of the field of Literature Related Discovery is available in (Kostoff, Block et al. 2009). I am not aware of any work that explicitly links LRD with heterogeneous data collections. Just as LRD has become possible with the digital indexing of citation data, we can presume that digital indexes of heterogeneous data collections might facilitate new forms of discovery algorithms.

Data Related Discovery

The nomenclature of "Data Related Discovery" is not in common use. In the legal profession, electronic discovery is a common term used to describe the process of electronically locating documents that are relevant to a particular case. I introduce the term here to refer to a particular subset of data mining and knowledge discovery and to contrast with Literature Related Discovery. As of November 22, 2009, a Google search for "Data Related Discovery" returns one hit and it is used analogously to electronic discovery on James Bowman's LinkedIn page.[15]

The April 17, 2009 issue of *Science* included a pair of articles on computational support for scientific discovery that reported on techniques that narrowed the gap from the analysis of large volumes of data to the generation of scientific theory (King, Rowland et al. 2009; Schmidt and Lipson 2009; Waltz and Buchanan 2009). Mass media coverage of these articles included headlines such as "Computer Program Self-Discovers Laws of Physics (Keim 2009)." In "Distilling Free-Form Natural Laws from Experimental Data (Schmidt and Lipson 2009)" one of the

[14] http://www.nlm.nih.gov/mesh
[15] http://www.linkedin.com/in/jamesbowman

experiments involved the input of motion tracking data recorded from a double-pendulum. “Without any prior knowledge about physics, kinematics, or geometry, the algorithm discovered Hamiltonians, Lagrangians, and other laws of geometric and momentum conservation (Schmidt and Lipson 2009).” It is notable that this work was funded by the NSF CreativeIT program. An NSF press release reports that the algorithms were actually developed for work on self-repairing robots and then the researchers realized their general applicability to a large data space. Using the terminology from literature retrieval, we can describe the algorithms used as search algorithms that covered the data space and produced minimally defined indexes to the data having maximal coverage of instances.

In “The Automation of Science” a robot “autonomously generated functional genomics hypotheses about the yeast *Saccharomyces cerevisiae* and experimentally tested these hypotheses by using laboratory automation (King, Rowland et al. 2009).” Again, an iterative search algorithm was used, however with the novel contribution that the robot was able to affect the physical world and generate new data points to define the search space during the exploration iterations. (Schmidt and Lipson 2009) closes with a description of the intended use of the work: “Scientists may use processes such as this to help focus on interesting phenomena more rapidly and to interpret their meaning.”

Thus, literature-related discovery and data-related discovery share commonalities in algorithmic design. Each use iterative data reduction and summarization to decompose a search space and each use pattern discovery to identify novel connections amongst elements of the decomposition. These classes of algorithms are generally explored in the fields of artificial intelligence and data mining (Hilderman and Hamilton 2001).

While literature related discovery and data related discovery each come from different historical traditions, they also share a common use case. Each provides utility by helping a researcher focus in on interesting elements within a large collection. “Michael Atherton, a cognitive scientist who recently predicted that computer intelligence would not soon supplant human artistic and scientific insight, said that the program [Schmidt and Lipson] ‘could be a great tool, in the same way visualization software is: It helps to generate perspectives that might not be intuitive’ (Keim 2009).”

The obvious difference between LRD and Data Related Discovery is the unit of analysis. However, the algorithms themselves are not tightly coupled to the raw input. Instead, they operate on indexes or surrogates of the input, particularly as further iterations generate reductions and summarizations of the full information space. Therefore, a combined information space of both literature and data has the potential to widen the scope of the discovery algorithms and therefore increase the potential for finding connections across more widely disparate elements. The methods described here include the construction of a combined information space within selected domains. The methods also include the development of algorithms to operation on the combined space.

Schmidt and Lipson explain that a problem with their current technique is that although the algorithms find descriptive and succinct equations, it is still a challenge to interpret the significance of those equations in the domain of study. They went on to say that this is a particularly difficult problem when analyzing bioinformatics data (Schmidt and Lipson 2009). A combined information space has the potential to address this problem. The high semantic density of a literature space can be used to contextualize patterns and unexplored elements of a data space when the two are correlated by a unified model.

Visual Analytics

In 2004, the United States Department of Homeland Security (DHS) chartered the National Visualization and Analytics Center (NVAC) at Pacific Northwest National Laboratory. Researchers from academia, industry, and government collaborated to develop a five-year research agenda and to define the grand challenges of the field. The results of this collaboration were published in "Illuminating the Path: The Research and Development Agenda for Visual Analytics (Thomas and Cook 2005)" one year later. The agenda was focused on addressing homeland security issues and intelligence analysis in particular. The grand challenges and research that the field produced are however generally applicable to any problems that require an understanding of complex data. Coincidentally, the same year that DHS and NVAC were publishing a book on the future of intelligence analysis the American Society for Information Science and Technology was publishing a book on its history. In "Covert and Overt: Recollecting and Connecting Intelligence Service and Information Science" the editors report:

> *"Originally, our intent was only to find some interesting speakers for a forthcoming professional conference. During the 2000 conference of the American Society for Information Science and Technology (ASIS&T), at a planning meeting for the Special Interest Group on History and Foundations of Information Science (SIG/HFIS), we undertook to arrange a session for the following year at which a panel of speakers would talk about their early backgrounds in intelligence work. It was already widely known, but rarely mentioned, that many of the people responsible for establishing the field of information science and for building ASIS&T into the leading professional*

> *association for the field had worked in intelligence agencies during World War II (Williams and Lipetz 2005)."*

The chapters of "Covert and Overt" are authored by individuals who provide personal accounts of their experience. Reading them one can see that a pervasive theme is the problem of huge volumes of records being generated and the consequent challenge in developing and maintained usable indexes. Many of the themes in the history of intelligence analysis and information science reappear in the Grand Challenge defined for Visual Analytics in "Illuminating the Path:"

> *"Grand Challenge: Enabling Profound Insights. One challenge underlies all of these objectives: the analysis of overwhelming amounts of disparate, conflicting, and dynamic information to identify and prevent emerging threats, protect our borders, and respond in the event of an attack or other disaster. This analysis process requires human judgment to make the best possible evaluation of incomplete, inconsistent, and potentially deceptive information in the face of rapidly changing situations (Thomas and Cook 2005, p.2)."*

While the problems of supporting human judgment with information are not new, modern increases in both volume (see section Information Overload) and in kind (see section Models of Scientific Communication) of available information are notable. It is recognized that visual representations of information can take advantage of aspects of human cognition in powerful ways. This has long been known in the field of cartography (MacEachren 1995). Leveraging these

cognitive capacities of the human visual system for information management with modern interactive computer graphics is therefore a promising path forward.

Lee S. Strickland, former intelligence office at the Central Intelligence Agency and professor at the University of Maryland, College of Information Studies, explicitly identifies the need for tools to integrate literature and data:

> *"Another key is addressing the volume of information – a veritable tsunami – and the need for tools. In short, the totality of information far exceeds the ability of any organization to effectively and completely analyze and render judgments. And there are several aspects to this issue. One is that textual information must be captured and must be retrievable. Another is that the textual information or structured data quickly outstrips the working capability of the mind to retain and this analyze.* ***Yet another is the necessity to integrate that unstructured text information with structured data.*** *These issues present a critical requirement: analytical software (tools) to work on the problems of entity and relationship extraction from texts as well as the analysis of the resulting data (e.g.,* ***the discovery of trends or links that are quite simply not obvious to the human analyst****)(Strickland 2005, p.164, emphasis added)."*

The Visual Analytics Science and Technology (VAST) Contests and Challenges were created to help support the development of new visual analytics tools by providing datasets with a hidden ground truth (Whiting, Cowley et al. 2006; Plaisant, Fekete et al. 2008). The use of a shared dataset helps to facilitate comparative evaluations of new tools and designs. The 2006 and 2007 VAST contest datasets made use of a corpus of textual data and award winning teams generally

leveraged entity extraction algorithms in combination with interactive visualizations (Görg, Liu et al. 2007; Stasko, Görg et al. 2007; Stasko, Gorg et al. 2008). In 2008 the format of the competition was changed with the introduction of mini-challenges (Grinstein, Plaisant et al. 2008). With this change, the mini-challenges were generally composed of structured data while the volume of textual data was reduced. The inclusion of image data remained small. Success in the Grand Challenge in 2008 required integrating multiple structured data sources and contextualizing the data by the narrative found in the textual elements. Winning entries generally made use of a graph data structure to integrate the heterogeneous sources while each structured data source was also given its own customized interactive visualization to support exploration (Chien, Tat et al. 2008; Payne, Solomon et al. 2008; Pellegrino, Pan et al. 2008).

Scenario Visualization

In "Scenario Visualization: An Evolutionary Account of Creative Problem Solving," Robert Arp contends that the human visual system has specifically evolved to allow humans to perform non-routine creative problem solving by making novel connections between previously unrelated information (Arp 2008).

> *"Unlike routine problem solving – which deals with associative connections within familiar perspectives – nonroutine creative problem solving entails an innovative ability to make connections between wholly unrelated perspectives or ideas (Arp 2008, p.9)."*

By viewing the human visual system as a hierarchical, modular system defined by information filtering and flow, Arp shows how humans chunk surrogates for raw images and then perform transformation operations on those surrogates to build novel connections. Arp defines scenario visualization as "a conscious activity whereby visual images are selected, integrated, and then

transformed and projected into visual scenarios for the purpose of solving problems in the environments one inhabits (Arp 2008, p.2)."

The essence of the discovery algorithms can be interpreted as a subset of the general process described by Arp. In the context of LRD, the chunks are coded as clusters of document surrogates (metadata) and the purpose of the algorithms is to identify connections between unrelated chunks that may be relevant to solving a problem in the world.

CHAPTER 3: STUDY DESIGN

Preliminary Studies

VAST Challenge

The National Visualization and Analytics Center (NVAC) have sponsored a Visual Analytics Science and Technology (VAST) Challenge annually since in 2006. The 2006 and 2007 events were referred to as the VAST Contest. The event was renamed to the VAST Challenge in 2008 when mini-challenges were introduced. "Its objectives were to provide the research community realistic tasks, scenarios, and data used in analytic work, to help visual analytics (VA) researchers evaluate their tools, and to improve and enrich interactive visualization evaluation methods and metrics (Plaisant, Grinstein et al. 2008)." The events were modeled off similar contests in other fields, such as the Text Retrieval Conference (TREC),[16] which plays an important role in evaluation for the field of information retrieval. We participated in the 2008 VAST Challenge in collaboration with the Pennsylvania State University as part of the North East Visualization and Analytics Center (NEVAC).[17] Our team entry was awarded a Grand Challenge Award for Data Integration (Pellegrino, Pan et al. 2008). Our experience in this event provided an opportunity to test methods for heterogeneous data integration. These experiences and methods served as a preliminary study and informed the study design described in the next section.

Mini-Challenges

The 2008 Challenge was organized into four separate mini-challenges. All of the data was synthetic. None of the scenarios or data records was from real-world collections. A description

[16] http://trec.nist.gov/
[17] http://www.geovista.psu.edu/NEVAC/

of the methods used for the generation of the synthetic data can be found in (Whiting, Cowley et al. 2006).

Each of the mini-challenges focused on a specific data collection. Additionally, the mini-challenges were embodied into a story that was told through textual documents. The Grand Challenge required integrating data from all of the mini-challenges into a complete story. The integrated set revealed plot elements that could not have been found within the mini-challenge sets alone. The team's approach to the challenge was to organize the team members into subgroups. Each subgroup focused on a specific mini-challenge. Additionally, weekly meetings were held for all members. During these meetings, the individuals and subgroups reported on their progress. The meetings also included a discussion of how individual findings might be combined to solve the Grand Challenge. Computer support was used for the meetings and collaborations. Adobe Connect[18] was used for same-time, different-place support in coordination with a conference call for the weekly meeting events. TWiki[19] was used for different-time, different-place collaboration. Email was also used extensively; however, email contents were not used to search for entities as the TWiki was. Custom interactive visualizations were built using Improvise (Weaver 2004) for each mini-challenge. Figure 3, Figure 4, Figure 5, and Figure 6 show screenshots from the custom tools that were included for the mini-challenges.

[18] http://www.adobe.com/products/adobeconnect.html
[19] http://twiki.org/

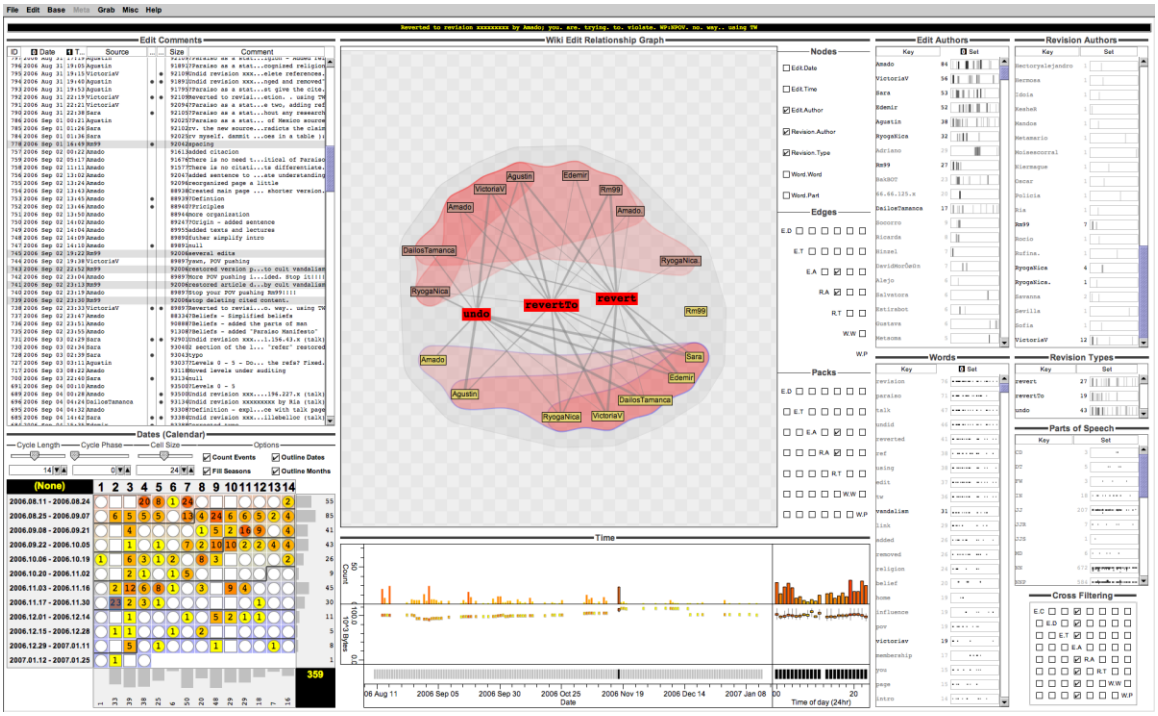


Figure 3: Custom Improvise visualization developed by Chris Weaver and the NEVAC team for analysis of the wiki collection.

Figure 3 shows a screenshot from the interactive visualization developed specifically for the wiki collection of the VAST 2008 mini-challenge dataset. This collection consisted of records in Wikipedia Page History format.[20] The full text of the historic versions of the pages was not available. Therefore, it was necessary to search for patterns within the revision log itself. For example, patterns of sequential revisions of one author by another could be interpreted as disagreement between the two authors on the topic of the page or section. The visualization tool was complemented by implementations of the analysis of controversy algorithms from (Brandes and Lerner 2008). Additional details on the specific techniques for the wiki collection are described in (Pan, Pellegrino et al. 2008).

[20] http://en.wikipedia.org/wiki/Page_history

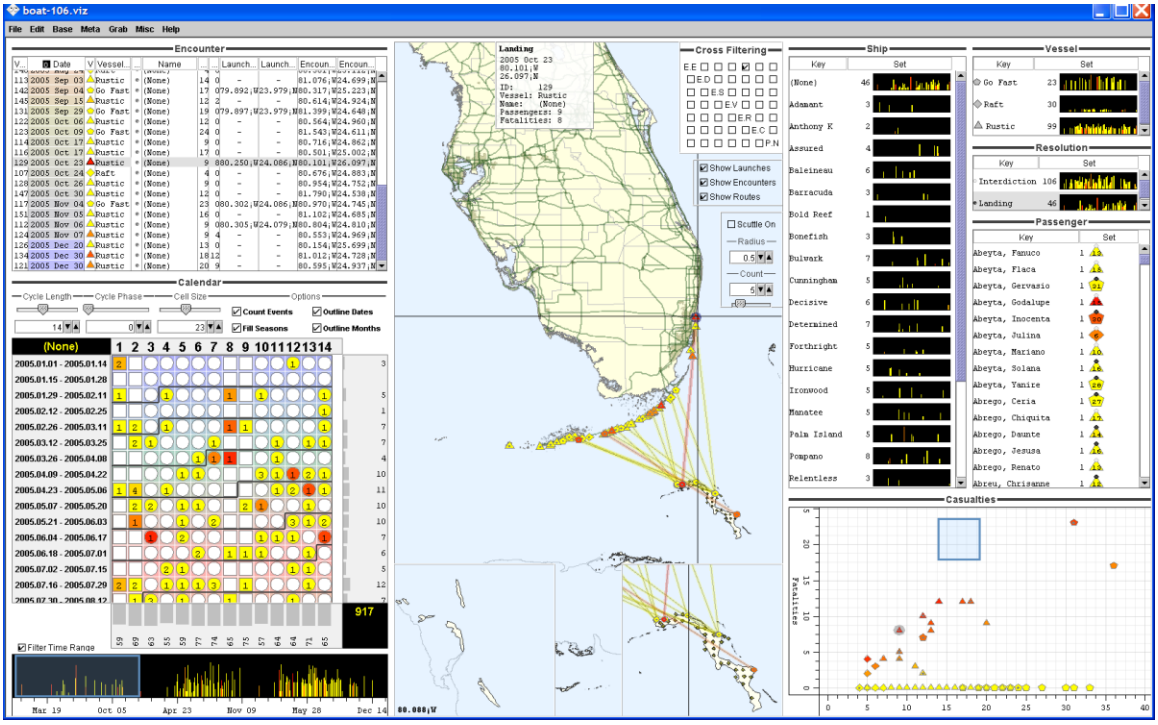


Figure 4: Custom Improvise visualization developed by Chris Weaver and the NEVAC team for analysis of the coast guard intercept collection.

Another mini-challenge data collection was formatted in XML. It consisted of synthetic data for US Coast Guard vessel interception events or vessel landing events. The fields of the XML nodes can be seen in the top-left component in Figure 4. The records also included a multi-valued field containing the vessel passenger list. In this mini-challenge, a fictitious island was imagined to be off the southern tip of Florida. Residents of the island were attempting to migrate to the United States. The storyline described a wet-foot, dry-foot policy for immigration. If a vessel landed in the United States then the passengers could stay in the country. If the vessel were intercepted at sea then the passengers would be returned to the island. Again, the mini-challenge involved the identification of patterns within the dataset. Relevant patterns included describing seasonal trends for vessel locations based on the latitude and longitude coordinates provided in the data. Other patterns included the detection of social networks based on passengers who frequently traveled together during attempts to reach land.

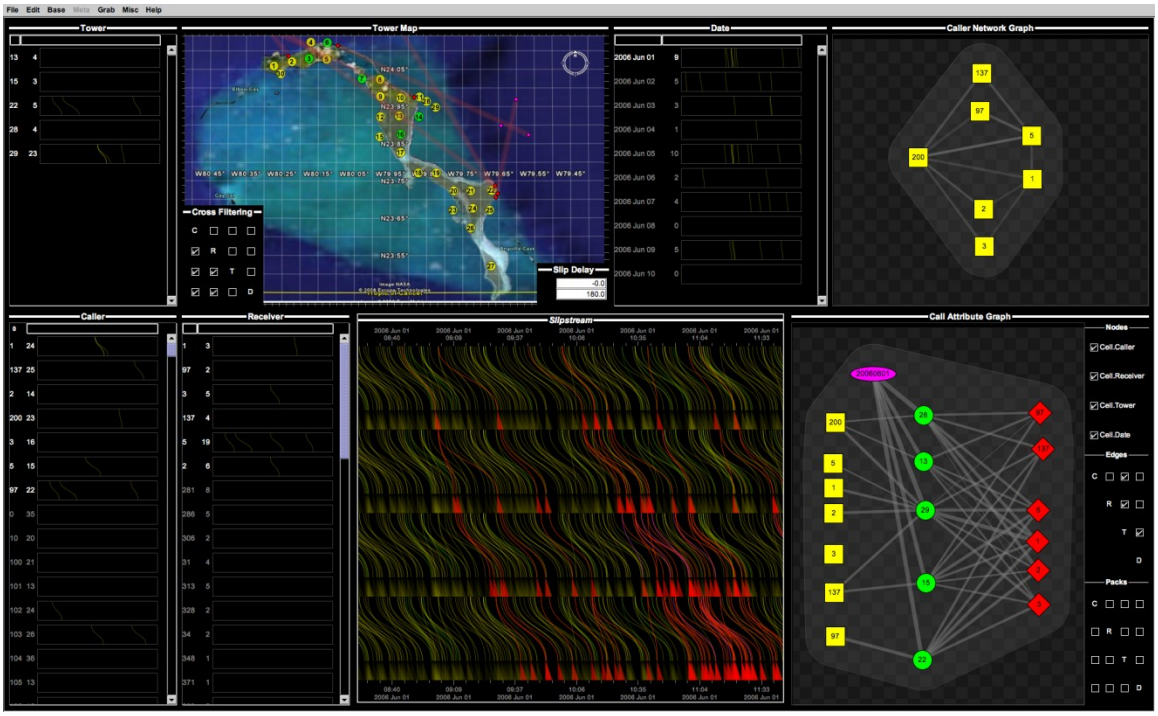


Figure 5: Custom Improvise visualization developed by Chris Weaver and the NEVAC team for analysis of the cell phone call collection.

The cell phone mini-challenge consisted of data is the common comma separated values format (CSV). The records included a timestamp, identifiers for the originating mobile phone, the target phone, and the cell tower that was used. The latitude and longitude for the fictitious cell towers was also provided in another file. Using the information a map could be constructed that showed the cell towers location on the fictitious island country. The interactive tool shown in Figure 5 was used to explore the data and to identify patterns. Patterns that could be found included calling activity across the island at certain times of day or calling activity in geographical regions over time. It was also possible to identify social networks by observing groups of phones that frequently called each other.

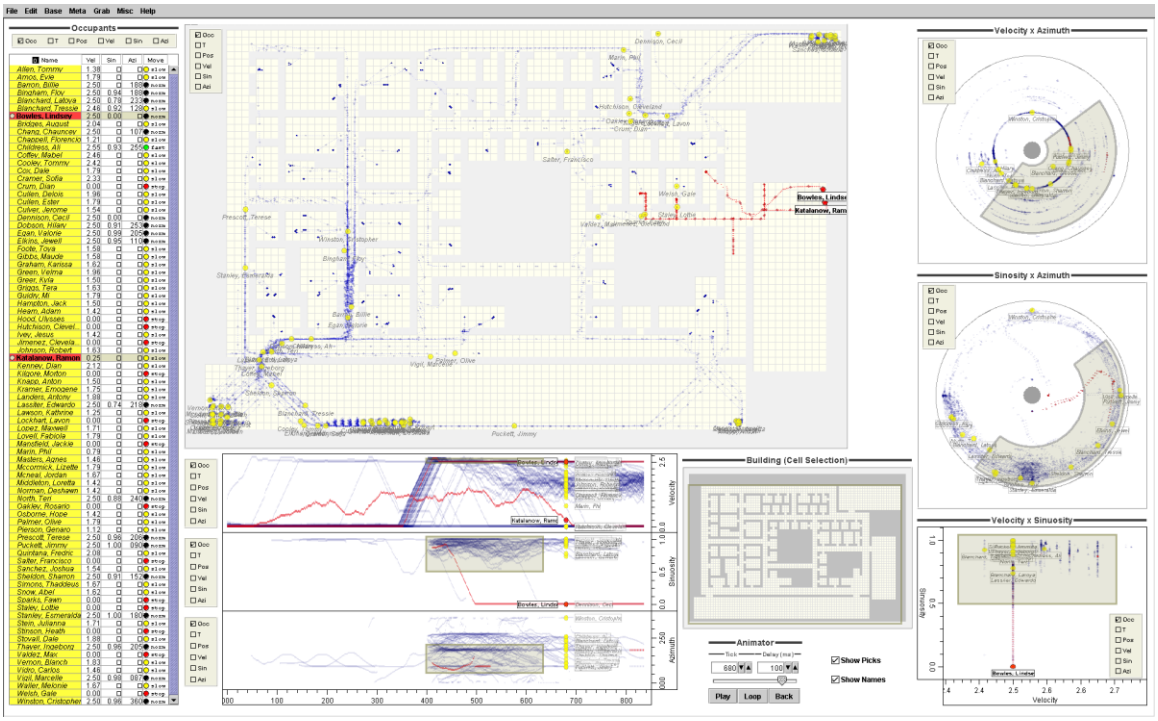


Figure 6: Custom Improvise visualization developed by Chris Weaver and the NEVAC team for analysis of the RFID movement collection.

The final mini-challenge dataset consisted of records of movement as tracked by RFID badges within a hospital building. In this story, the RFID badges were assigned to hospital staff and to visitors. Figure 6 shows a screenshot from the interactive visualization tool that was developed to analyze this collection. Animation was used to show the movements of badges. The story for this dataset described a bomb being set off in the hospital. Patterns that could be observed included the movements of personnel during the normal time before the bombing event and the evacuation that occurred after. A goal in this mini-challenge was to find patterns that could be used to identify suspects who may have been involved in the bombing, based on their behavior before, during, and after the event.

Grand Challenge

The Grand Challenge required constructing a storyline out of the datasets from the individual mini-challenges. Our full submission for the Grand Challenge has been archived by the National Institute of Standard and Technology (Pellegrino, Chen et al. 2008). The materials for the event included additional text including a synthetic "Paraiso Manifesto" Wikipedia article describing a religious organization on the fictitious island. These additional text materials were treated as a fifth dataset for the Grand Challenge. Although weekly meetings were held with attendance by all team members, we were still in search of a method for making sense of all of the mini-challenges and figuring out how everything fit into a main plot. One technique for helping to systematically generate hypotheses is to search for patterns in entity graphs created from the data (Pellegrino and Chen 2008). By treating the hypotheses developed by the sub-teams working on the mini-challenges as data, we were able to construct a graph that including records from all of the mini-challenges along with the higher-level ideas about those collections that had been recorded in the team's wiki.

> *"Integration of the data and findings was done by using an associative network as the fundamental data structure. This provided the greatest degree of abstraction while preserving the critical connectedness between the different types of data. A transform was created for each of the four mini-challenge data sets. The transforms created nodes in the network to represent their connectedness. The hypotheses and assumptions captured in the Wiki were represented as derived nodes and edges in the network. These constructs helped to assign a higher-level meaning to the data making the model a semantic network rather than simply a set of*

associations. By combining higher-level constructs such as hypotheses with the raw data, analysis results became more useful. (Pellegrino, Pan et al. 2008)"

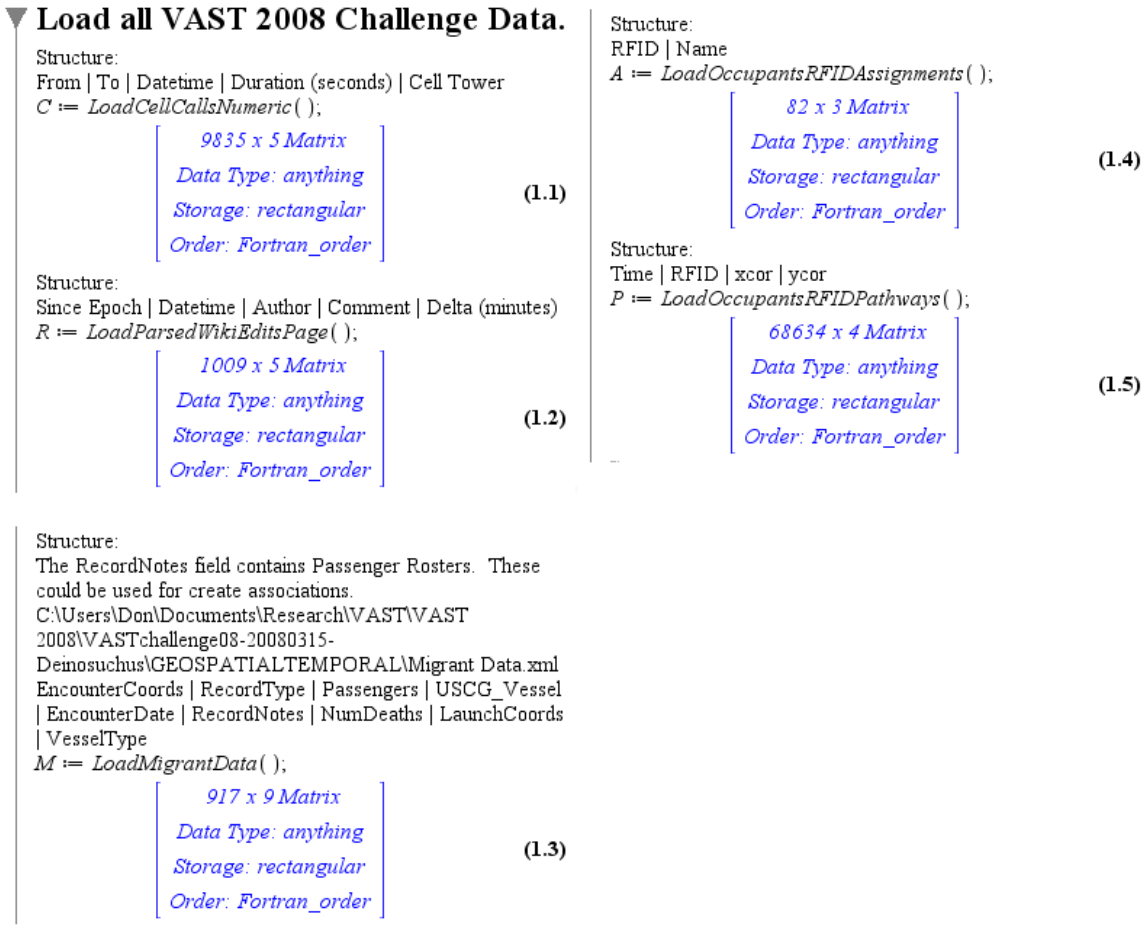


Figure 7: All of the mini-challenge data collections were loaded into a single Maple worksheet. (Pellegrino, Chen et al. 2008, Figure 1)

Figure 7 shows the Maple[21] implementation used to load all of the separate mini-challenge datasets into memory within a single Maple worksheet. The procedures "LoadCellCallsNumeric," "LoadParsedWikiEditsPage," "LoadMigrantData," "LoadOccupantsRFIDAssignments," and

[21] http://www.maplesoft.com/products/Maple/index.aspx

"LoadOccupantsRFIDPathways" were custom built to ingest the unique formats from each of these collections into a standard matrix structure.

```
# Hypotheses
H_casualties := vertex;
fprintf(fd, "%d \"H Casualties\"\n", vertex);
vertex := vertex + 1 :

H_suspects := vertex :
fprintf(fd, "%d \"H Suspects\"\n", vertex) :
vertex := vertex + 1 :
```

```
# Results of the Evacuation Mini-Challenge.
for i from 1 to 82 do
 # Add the Casualties Hypothesis.
 if evalb(A[i, 1]in {18, 19, 56, 36, 76, 50, 39, 78, 65, 60, 47, 69}) then
  fprintf(fd, "%d %d\n", AID[i] + 1, H_casualties);
  fprintf(fd, "%d %d\n", H_casualties, AID[i] + 1);
 end if;

 # Add the Suspects Hypothesis.
 if evalb(A[i, 1]in {21, 1, 29, 44, 56}) then
  fprintf(fd, "%d %d\n", AID[i] + 1, H_suspects);
  fprintf(fd, "%d %d\n", H_suspects, AID[i] + 1);
 end if;
end do:
```

Figure 8: "Modeling the evacuation mini-challenge hypotheses in an associative network (Pellegrino, Chen et al. 2008, Figure 7)."

An associative network was instantiated from the tuples of the mini-challenge data collections. To establish meaningful linkages amongst records within the challenge data, we encoded the hypotheses that had been formed by the sub-teams. The hypotheses themselves were instantiated as nodes in the graph. The entities or records references by those hypotheses were connected as edges between the hypothesis nodes and the evidence nodes. Figure 8 shows the encoding used to represent the hypothesis of bombing suspects. RFID tags 21, 1, 29, 44, and 56 were suspected of being involved in the bombing. This hypothesis had been developed by the sub-team working with the interactive visualization of the RFID data and the hospital bombing subplot. The hypothesis had been captured in the team wiki. The Maple source code to create the graph elements for the hypothesis had to be manually written after the hypothesis had been identified and documented as text.

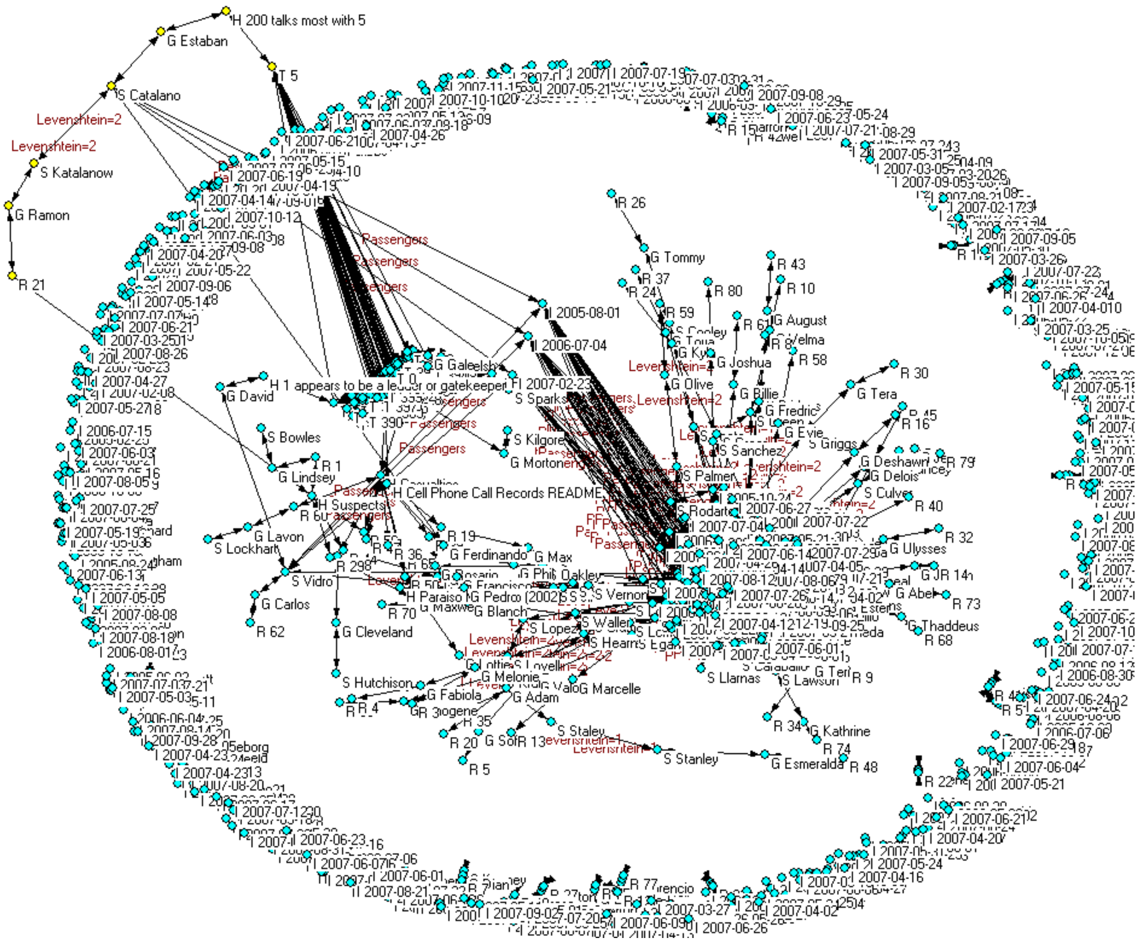


Figure 9: Graph representation of data and hypotheses (Pellegrino, Chen et al. 2008, Figure 8).

Figure 9 shows the graph created by combining the records from all of the mini-challenges along with the node and edge encodings for the multiple hypotheses. Maple was used to create the definition of the graph however, Pajek (de Nooy, Mrvar et al. 2005) was used for the layout, projection and interaction with the resultant graph. A breakthrough connection was established with the observation that the surnames “Katalanow” and “Catalano” sound similar even though they are spelled differently. This background material for the RFID data indicated that RFID tags were assigned to visitors to the hospital. Within this context, it was possible to envision a data entry error for the spelling of a visitor’s last name. Once recognized through interaction exploration of the data visually, we were then able to encode logic into the graph generation

algorithm. An edge was instantiated between any two nodes that had text with a Levenshtein similarity score within a threshold. The addition of this algorithm allowed for the systematic exploration of relationships of that type, of name misspellings, across fields from different data sets that had originally been encoded with different formats and different semantics. Using shortest path analysis with Pajek it was then possible to identify links between hypotheses and data for which no one on the team had previously seen connections. These connections between high-level hypotheses and low-level data provided useful insights to the team. Figure 9 shows a path from a RFID tag five from a hospital in Florida during the bombing, to mobile telephone 5 on the island by way of data records and a hypothesis about calling behavior on the island.

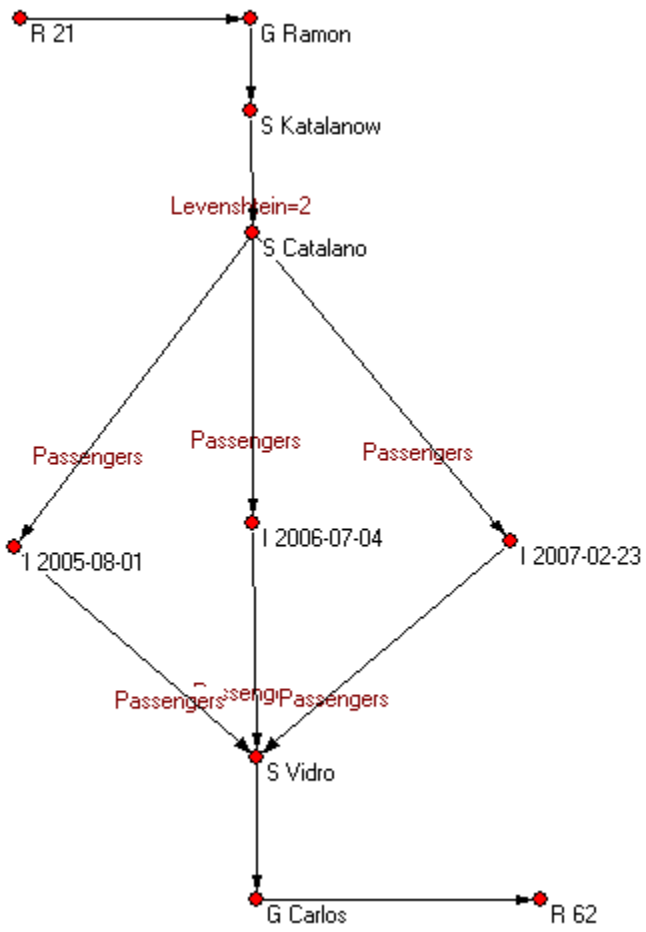


Figure 10: "Path from RFID 21 to RFID 62 (Pellegrino, Chen et al. 2008, Figure 10)."

Another example of an insight found using the method is described by Figure 10. This figure shows a path within the graph from RFID 21 to RFID 62. Carlos Vidro and Ramon Catalano had been assigned these RFID tags and they had been passengers together according to three separate vessel records in the Coast Guard collection. The identification of this path provided insight to the RFID sub-team. With this information, they could then look for patterns in movement data for RFID 21 and RFID 62 that they may have missed earlier.

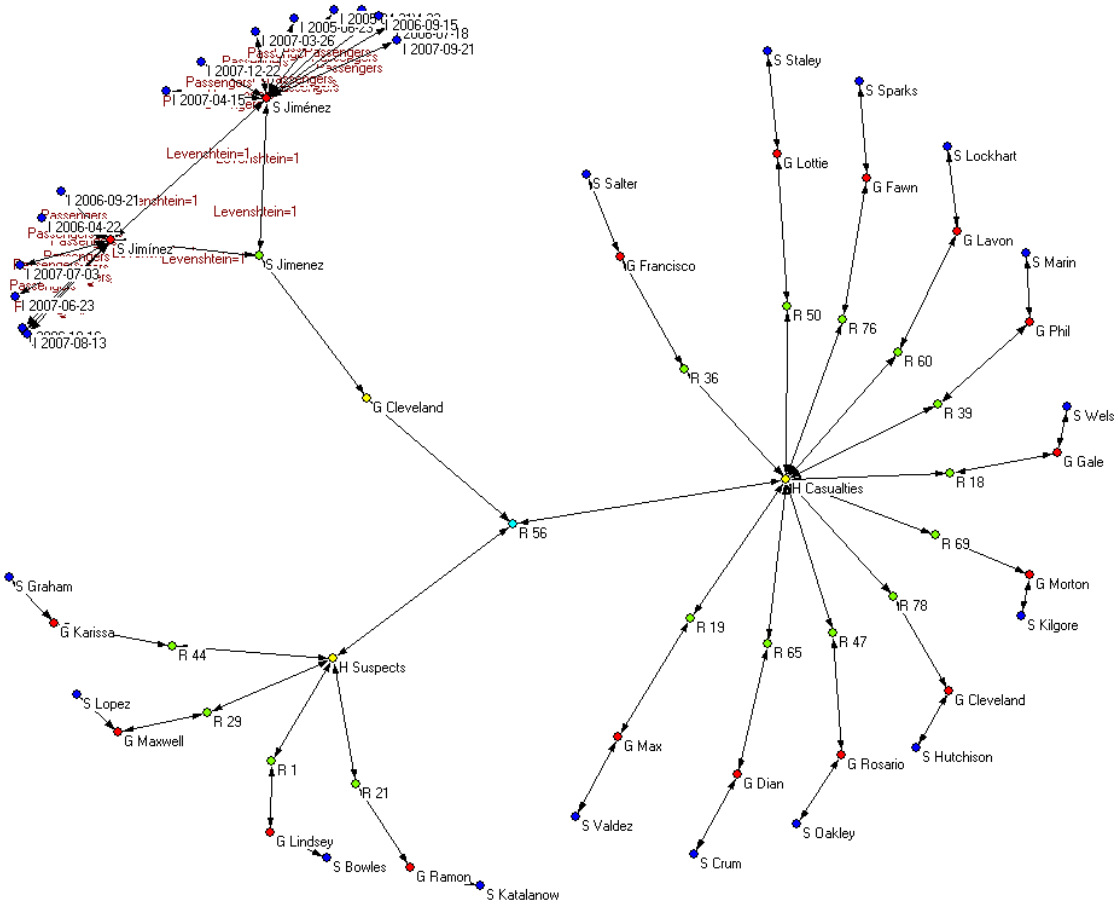


Figure 11: "k-Neighbors within 4 of RFID 56 (Pellegrino, Chen et al. 2008, Figure 11)."

Just as useful as finding new paths for exploration is the process of identifying conflicts or misunderstandings. Figure 11 shows two conflicting hypotheses. Cleveland Jimenez, who had been assigned RFID 56, had been hypothesized to be both a suspect and a casualty. A k-nearest neighbor algorithm can be used to analyze this discrepancy. Focusing on the conflicting evidence node of "R 56 (RFID number 56)," the 4-nearest neighbors in the full graph are shown. This provides a visual contextualization of the evidence within the full domain of encoded knowledge. The method was useful for considering whether the individual might have been a suicide bomber or whether one of the suspect or casualty hypotheses was incorrect. By

providing the context of connected evidence, the team was able to identify paths for further exploration of this conflict.

Learnings

Our experience in the VAST Challenge 2008 provided many valuable lessons and opportunities for further study. The screen sharing and wiki tools we selected for collaborative support are commonly used. The divide-and-conquer approach we used to organize the team is also typical. Methods for systematically supporting the discovery of critical linkages were not available as features in existing tools. The key insights came from links that traversed both evidence data and higher-level hypotheses, which had been formulated about the data by individual or groups. While the input data was well structured, our hypotheses about the relationship of data to the story were captured as text in the wiki. A fundamental problem had been recognizing how the data, and our ideas about the data, were connected. The method of instantiating nodes and edges for each of the hypotheses was very successful for generating insights. However, this method required substantial manual effort. It was necessary for someone to read all of the hypotheses that had been recorded in the wiki. Those narrative structures then had to be encoded as nodes and edges in the graph. Further, Maple programming skills were needed for embedding the encodings in a way that was compatible with the graph that was built automatically from the input data. This dependency on manual effort by an individual with a sophisticated and specific skill limited the generalizability of the approach.

A natural question is whether the methods that had been successful for us in the VAST Challenge could be applied in other domains. Additionally, the methods could be more broadly accessible if the data and ideas about the data could be processed automatically, or at least

semi-automatically. The studies described in the next section describe the approach taken to address these questions.

Study Design

Specific Aims

Answers to the following research questions will increase our understanding of the relationship between information stored as data and information contained in the literature. This understanding will be applied to the purpose of systematically supporting the creative process of making novel connections.

RQ_1	What are the graph theoretic properties of graphs that are created from the combination of collections of literature with databases of data?
RQ_2	Which connections become available in the combined information spaces that are not available in either source individually?
RQ_3	Can novelty detection algorithms systematically use these graphs to identify connections that experts will find both novel and useful?

Table 2: Research Questions

Methodology

To answer the research questions in Table 2, three series of experiments have been carried out. In each series, a domain-specific collection was selected for study. The dashed outline drawn around Figure 2 indicates Søndergaard's observation that different epistemologies in a given domain will emphasize different knowledge sources (Søndergaard, Andersen et al. 2003). The method accounts for this by allowing for the adaptation of the implementations to the domain-specific emphases that are important to researchers. Within each series, the research questions

are qualified by the domain. The methodologies are also designed to conform to a general pattern while being adapted to the specifics of the information available in the domain.

In the first series, a large collection of protein sequence records from studies of the influenza virus was connected. This series is described in “CHAPTER 4: Influenza Protein Sequence Analysis.” We this series we identified engineering challenges of creating and visualizing large graphs of data. A particularly challenging aspect of this work was the implementation of a prototype system that could present an overview of all of the data while also providing interactivity. We describe our solution to this problem and some of the advantages our solution has over existing systems.

In the second series, electronic laboratory notebooks from an Open Notebook Science system were connected with molecular properties from a large, crowd-sourced molecular structure database. This series is described in “CHAPTER 5: Open Notebook Science.” We developed a method for generating an overview representation of the data. We were also able to use the overview to identify a significant relationship that both provided critical insight to researchers and was difficult to find using existing methods.

The third series connected candidate drug molecules developed by Pfizer with corporate records for the research projects that synthesized the molecules. The third series is described in the “CHAPTER 6: Pfizer Drug Discovery Projects.”

These three series share the commonality that they all deal with molecular information. The methods and prototype systems developed within each series are specialized to the scientific problems of each domain. The corollary to this specialization is that the methods are not directly generalizable to other collections. The use of multiple series allows us to address this limitation

through triangulation of the results. By doing so, we can make inferences about the generalizability of the findings.

To answer research question (RQ_1), graphs were instantiated from the selected datasets and collections. Vertices in the graphs were created from both database records and literature collection records. Edges in the graphs were created between the vertices representing database records and the vertices representing articles. Edges were also created among the database records to each other and among the articles to each other. These graphs were then analyzed using in terms of their graph-theoretic properties. Example graph-theoretic properties include measures of topology, degree centrality, clustering and repeating sub-graphs. Many tools are available for calculating these measures. For a listing of relevant tools, see "APPENDIX A: Graph Visualization Tools."

With the graphs in place, answers to research question (RQ_2) were found by examining the connections that could be built and then subtracting the sets of connections that could have been built using graphs derived from the either the database records or the literature records alone. This increased our understanding of the information that is added by connecting these two information spaces.

Interviews with experts were used to answer research question (RQ_3). The graph theoretic properties of the connections that are unique to the combined information space were joined with user assessment of their meaning in the domain. These assessments were used to define heuristics. Heuristic strategies can be used to develop algorithms that systematically identify connections that have a high likelihood of providing novel and useful information to a domain expert. Desirable information that cannot be found algorithmically was used to define the limits of this approach. The inability to find novel or useful connections given an exhaustive search

was used to identify algorithms that have limited utility and that can be excluded from further investigation.

CHAPTER 4: INFLUENZA PROTEIN SEQUENCE ANALYSIS STUDY

Project Description

The influenza protein sequence analysis project is reported in detail in (Pellegrino and Chen 2011). The abstract for the paper reads:

> *"This paper introduces a new method for creating an interactive sequence similarity map of all known influenza virus protein sequences and integrating the map with existing general purpose analytical tools. The NCBI data model was designed to provide a high degree of interconnectedness amongst data objects. Substantial and continuous increase in data volume has led to a large and highly connected information space. Researchers seeking to explore this space are challenged to identify a starting point. They often choose data that is popular in the literature. Reference in the literature follow a power law distribution and popular data points may bias explorers toward paths that lead only to a dead-end of what is already known. To help discover the unexpected we developed an interactive visual analytics system to map the information space of influenza protein sequence data. The design is motivated by the needs of eScience researchers. (Pellegrino and Chen 2011)"*

The preliminary study of the VAST 2008 Challenge had demonstrated the feasibility of instantiating a graph from all of the records of interest in the relevant data collections. Figure 7 shows the load routines used for the VAST collections and the resultant matrices. These are summarized in Table 3.

Collection	Records	Dimensions	Nodes (Records × Dimensions)
Cell Calls	9835	5	49175
Wiki Edits	1009	5	5045
Migrant Data	917	9	8253
RFID Assignments	82	3	246
RFID Pathways	68634	4	274536
	80477	**26**	**337255**

Table 3: Summary of collection size from the 2008 VAST Challenge set.

One of the questions resulting from the VAST experience was whether the data-structure of a single associative network graph would also work when dealing with a substantial real-world scientific dataset. The maximally sized graph for the VAST Challenge collections could have included 337,255 nodes. In practice, the "RFID Pathways" collection was not included in the graph, yielding a substantially smaller working graph maximum of 62,719. These numbers exclude the hypotheses nodes that were manually added, however these were at most in the dozens. Working with the smaller graph made the techniques used in the VAST Challenge feasible. Investigations into including the "RFID Pathways" collection were not performed. Therefore scaling up the methods remained an open question.

During 2009, influenza research took on increased social importance with the outbreak of a novel strain that lead to a pandemic in humans (Cohen 2009). The National Center for Biotechnology Information (NCBI) at the US National Institutes for Health (NIH) provides a centralized resource for influenza sequence data (Bao, Bolotov et al. 2008). Sequences from the novel strain were available publicly through NCBI within a month of initial epidemiological detection (Cohen 2009). This public collection of data in combination with the substantial increase in research on influenza provided an excellent opportunity to explore the applicability of methods from the VAST Grand Challenge data integration to a real-world collection in another domain.

Summary of Findings

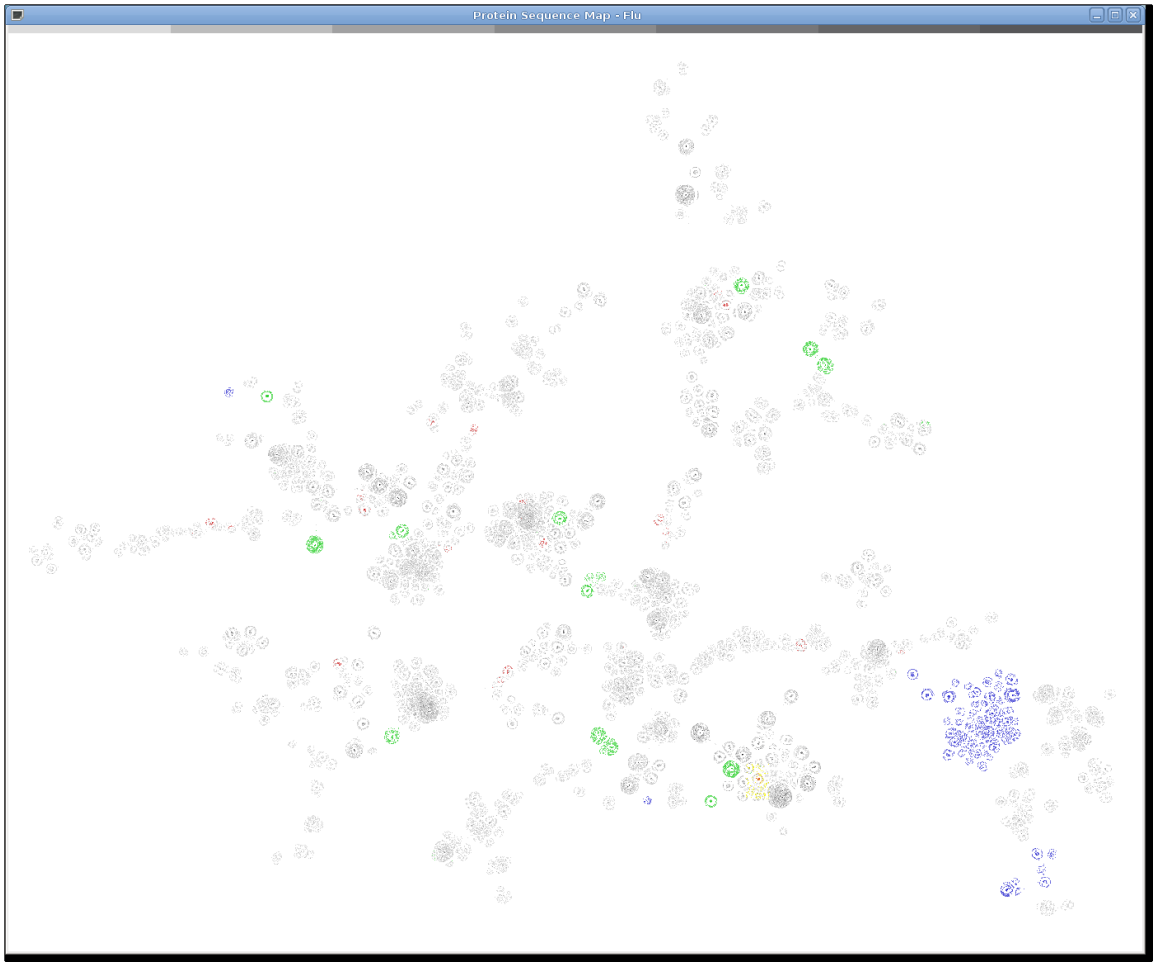


Figure 12: "Influenza virus protein sequence similarity map. 114,996 influenza virus protein sequence records from NCBI as of August 7, 2009 are shown. Sequences from the 2009 H1N1 Swine Flu pandemic are colored green. Sequences from the 1918 H1N1, 1957 H2N2, and 1968 H3N2 deadly human pandemics are colored red. Sequences that code for the PB1-F2 protein known to cause virulence in humans are colored blue. (Pellegrino and Chen 2011, Figure 3)"

Figure 12 shows a projection of 114,996 protein sequence records from NCBI as of August 7, 2009. Nodes were instantiated for each record in the collection. Edges were instantiated between records that met a BLASTP similarity score threshold. Supercomputers were used to perform the 13 billion pairwise comparisons (114,996 × 114,996) and to calculate the two-dimensional projection using the Large Graph Layout (LGL) algorithm from (Adai, Date et al. 2004).

The size of the resulting graph exceeded the limits of available tools to perform real-time interactive analyses. Operations such as scaling, rotating, zooming and selection were problematic when performed on the entire graph. A custom-built C program was written to affect greater hardware control. The OpenGL API was utilized to create node and edge geometry in dedicated graphics hardware memory and to allow for full graphics processing unit (GPU) acceleration of these operations. The full source code for the interactive system we build is available online.[22] Figure 12 is a screenshot from this system running on a Linux workstation. Figure 13 is another screenshot from the same system demonstrating an interactive zoom and selection performed in coordination with an external statistical tool.

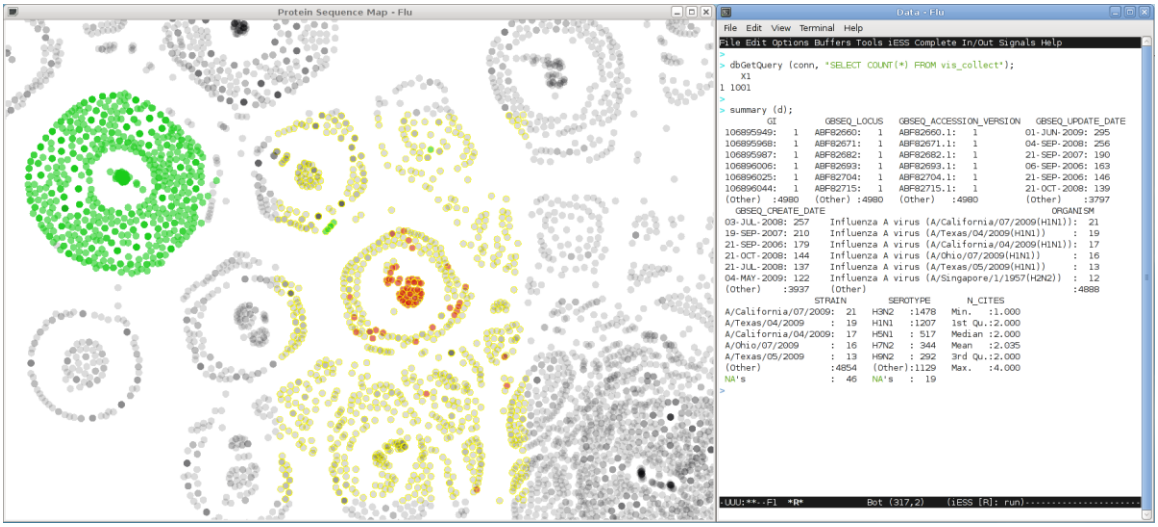


Figure 13: "Interactive influenza virus protein sequence similarity map (left, custom tool) integrated with general purpose analytical tools (right, Emacs and the R program for statistics). A set of 1001 sequence records are selected from a zoom region. The full map, shown in Figure 3, represents 114,996 sequence records. (Pellegrino and Chen 2011, Figure 4 - note reference to Figure 3 is relative to the original paper)"

Using the interactive visualization to explore the data researchers can see a selection of records in context with the entire collection. In Figure 13, the set of green circles in the center of the map on the left side of the figure represent sequences from the 2009 swine flu pandemic. It can

[22] http://cluster.ischool.drexel.edu/~st96wym4/flumap/

be observed that these mutations have more in common with the red sequences from more deadly human pandemics then those deadly pandemics had with each other. This method could be of potential use for disease monitoring and epidemiology.

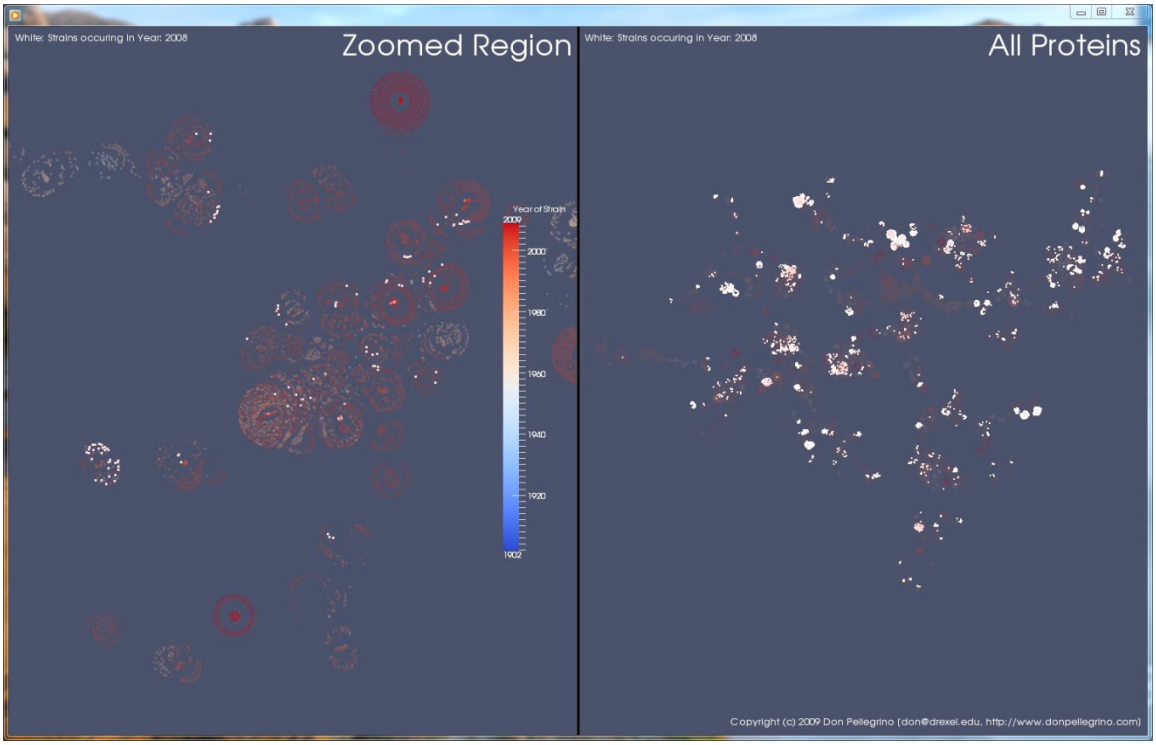


Figure 14: Sequence records registered with NCBI in 2008 versus all records.

The graph data and LGL projection were also used in an animation to explore the temporal dimension of the collection. Figure 14 shows sequence records registered with NCBI in 2008. During 2008 and prior years the sequence diversity of records registered for the given year spanned the entire breath of the diversity of the collection.

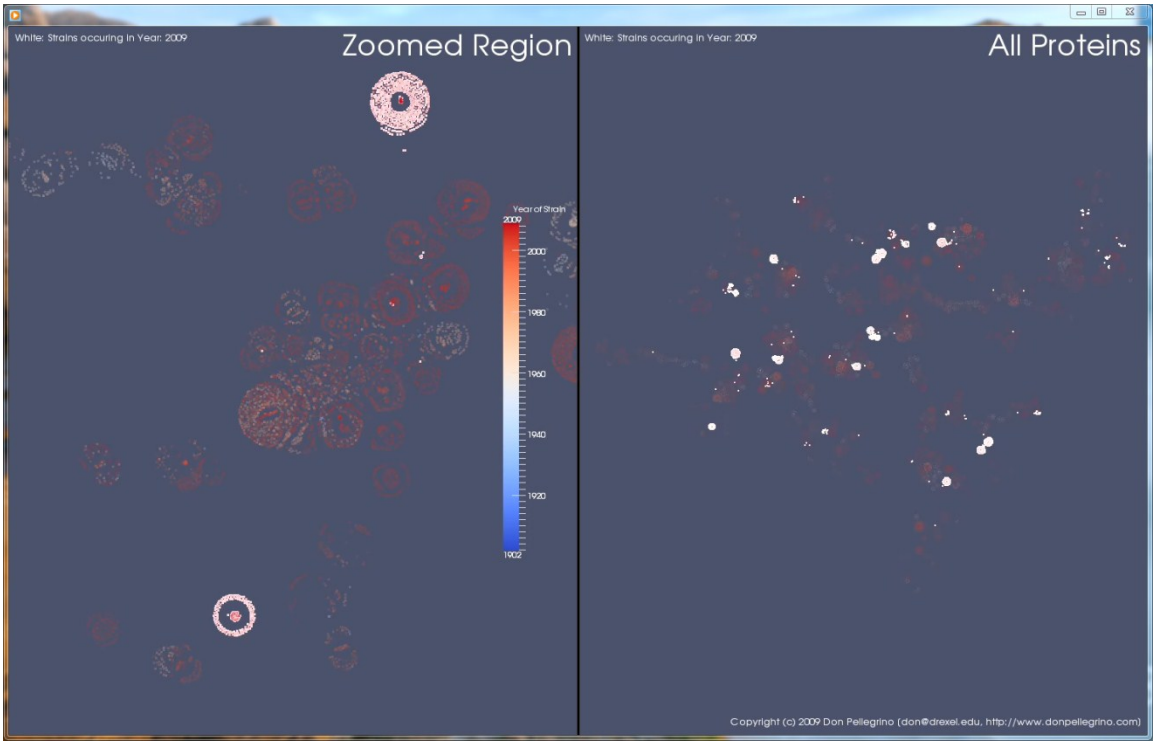


Figure 15: Sequence records registered with NCBI in 2009 versus all records.

Figure 15 shows the sequence records registered during the pandemic year. In this case, a macroscopic property of the year can be observed that cannot be represented with alternative systems. During the pandemic year, the diversity of the collection dropped drastically, with records appearing in tight clusters. The white nodes in the zoom region shown in the left frame of Figure 15 show this effect. The records uploaded to NCBI do not constitute a census of influenza virions. The sampling is not systematic. Therefore, at least two explanations are possible. It may be that researchers only studied the pandemic strain at that time and therefore the collection is biased toward that strain. Alternatively, it could be that the pandemic strain was so dominant that it led to the extinction of other strains.

Lessons Learned

The Influenza Protein Sequence Study demonstrated that the graph data structure could be scaled from the small collections studied in the 2008 VAST Challenge to the real-world collection

of protein sequence data. During the course of this study, limitations on interactive exploration of large graphs with current tools were identified. These limitations were overcome with the construction of a new system. The new system made use of C programming and the OpenGL API for better hardware control. This system was then used for an analysis of the graph. Together these provided a new overview that provided novel insight on the dynamics of the collection. They also provided a new means for identifying interesting members of the collection that would be more difficult to identify using the form and filter methods of existing web form-based systems.

CHAPTER 5: OPEN NOTEBOOK SCIENCE STUDY

Project Description

"Open Notebook Science is the practice of making the entire primary record of a research project publicly available online as it is recorded (Wikipedia contributors 2009)." For additional background on Open Notebook Science see the related section in "CHAPTER 2: Literature Review." A fundamental advantage of Open Notebook Science (ONS) is that it provides a complete chain of provenance from research conclusions to experimental activities. The activities and observations of a researcher are recorded shortly after they occur. Although many scientific fields have adopted the use of a laboratory notebook, only a few laboratories have begun to publish the contents of those notebooks online while they are written. One such laboratory is the Bradley Laboratory at Drexel University.

In the Literature Review section, we saw a path from the challenges of indexing intelligence data during World War II after the cracking of the Enigma, to indexing the literature by citation references, to today's challenge of indexing the hyperlink structure of the World Wide Web. In parallel to developments in indexing were changes to the kind of artifacts being published, as described by the UNISIST models of 1971 and 2003. The Open Notebook Science practice creates new kinds of artifacts and in large volumes. In addition to the laboratory notebook text, structured databases of data are also being produced and published publicly. This creates an opportunity to revisit indexing methods.

Cartography provides a useful analogy for the evaluation of indexes. A core concern for a cartographer is deciding which pieces and concepts from the natural world should be represented in the map. Mapmakers are faced with many design criteria, such as the selection of appropriate colors, relief, scale, and symbology. A critique of a map may address each of

these design decisions. In “How Maps Work,” Alan MacEachren suggests that maps should be evaluated by how well they serve their intended purpose (MacEachren 1995). Thus, each design decision can be considered for its suitability to the user of the map. Similarly, the multiple design considerations for an information index must be balanced and selected based on their suitability for a purpose. In this project, we have selected identification of connections that have the potential to provide support for discovery as the purpose of the index. The users of the index are chemists who are trying to answer the question, “What reaction should I perform next?”

Since they are by definition open, many of the artifacts in Open Notebook Science are already indexed by search engines, such as Google, which may be based on an approach derivative of PageRank (Page 2001). The most common interface to such search engines is a text box allowing the user to enter a query string. A design-time problem in the creative discovery process performed by a chemist is the formulation of the next research question. One challenge chemists are faced with is identifying relevant reactions that have already been performed. The details of those reactions support the decision of whether to run a new reaction or to rely on existing data.

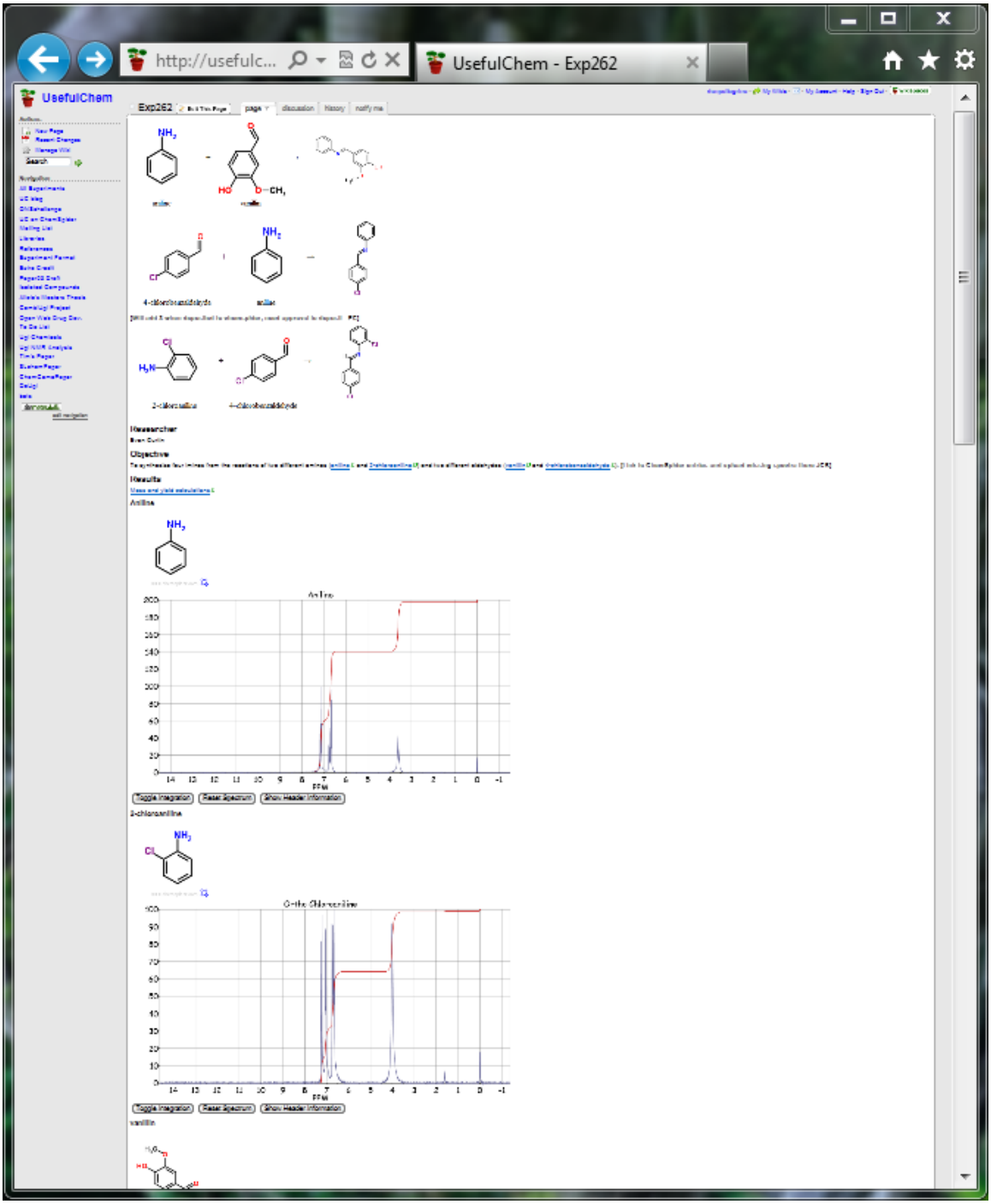


Figure 16: UsefulChem Experiment 262 Notebook Entry by Evan Curtin – part 1.

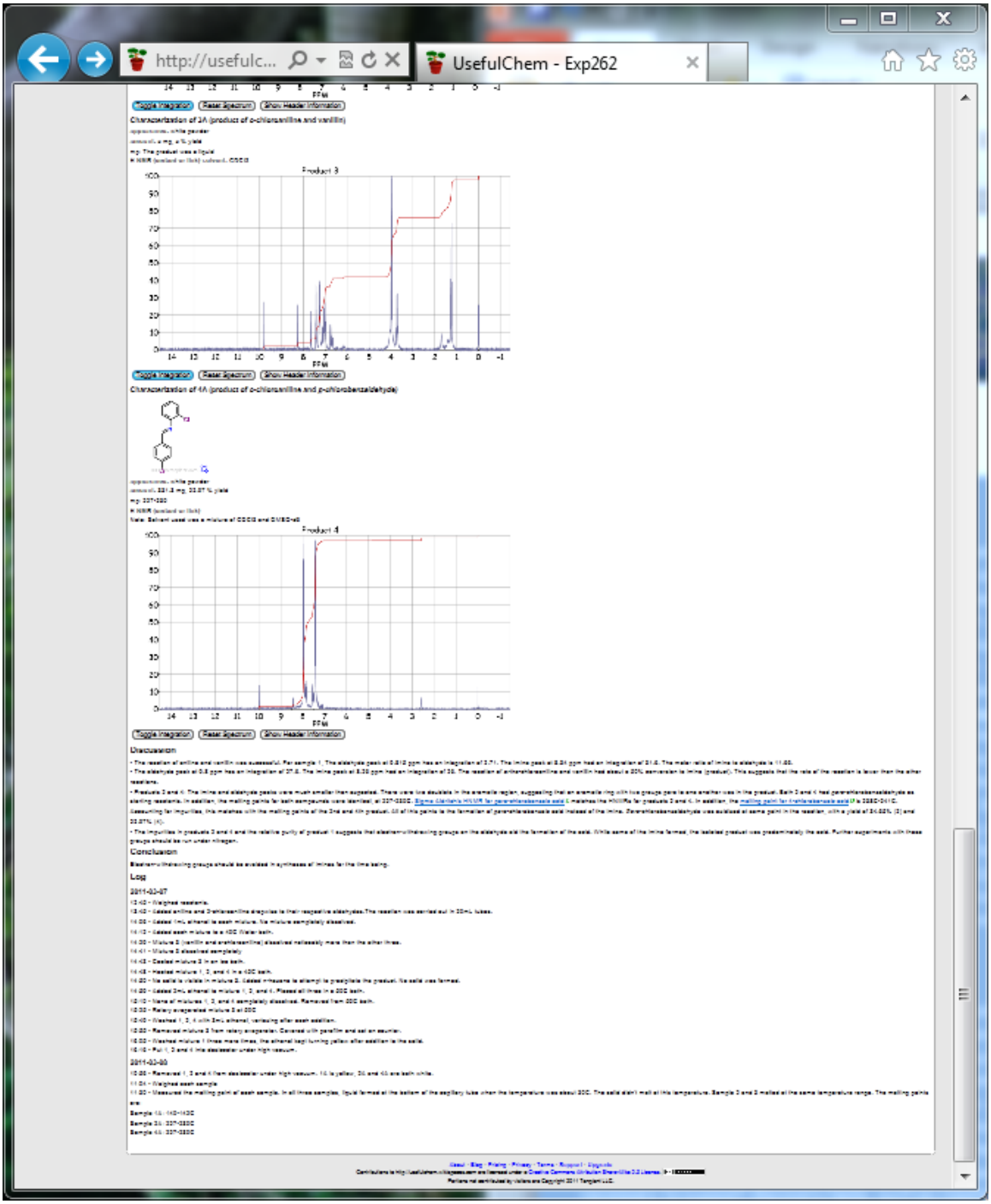


Figure 17: UsefulChem Experiment 262 Notebook Entry by Evan Curtin – part 2.

Figure 16 and Figure 17 show an example notebook page from the UsefulChem site. UsefulChem Experiment 262 by researcher Evan Curtin shows an entry following a typical structure. An Objective section describes the goal of the experiment. A Procedure section describes how the

experiment was carried out. Content includes text as well as structures from ChemSpider and NMR output. The Discussion and Conclusion sections describe the experiment and its interpretation. The Log section describes actions taken and observations made. These examples were also used in (Pellegrino, Bradley et al. 2011).

Summary of Findings

A Model of Open Notebook Science for Organic Chemistry

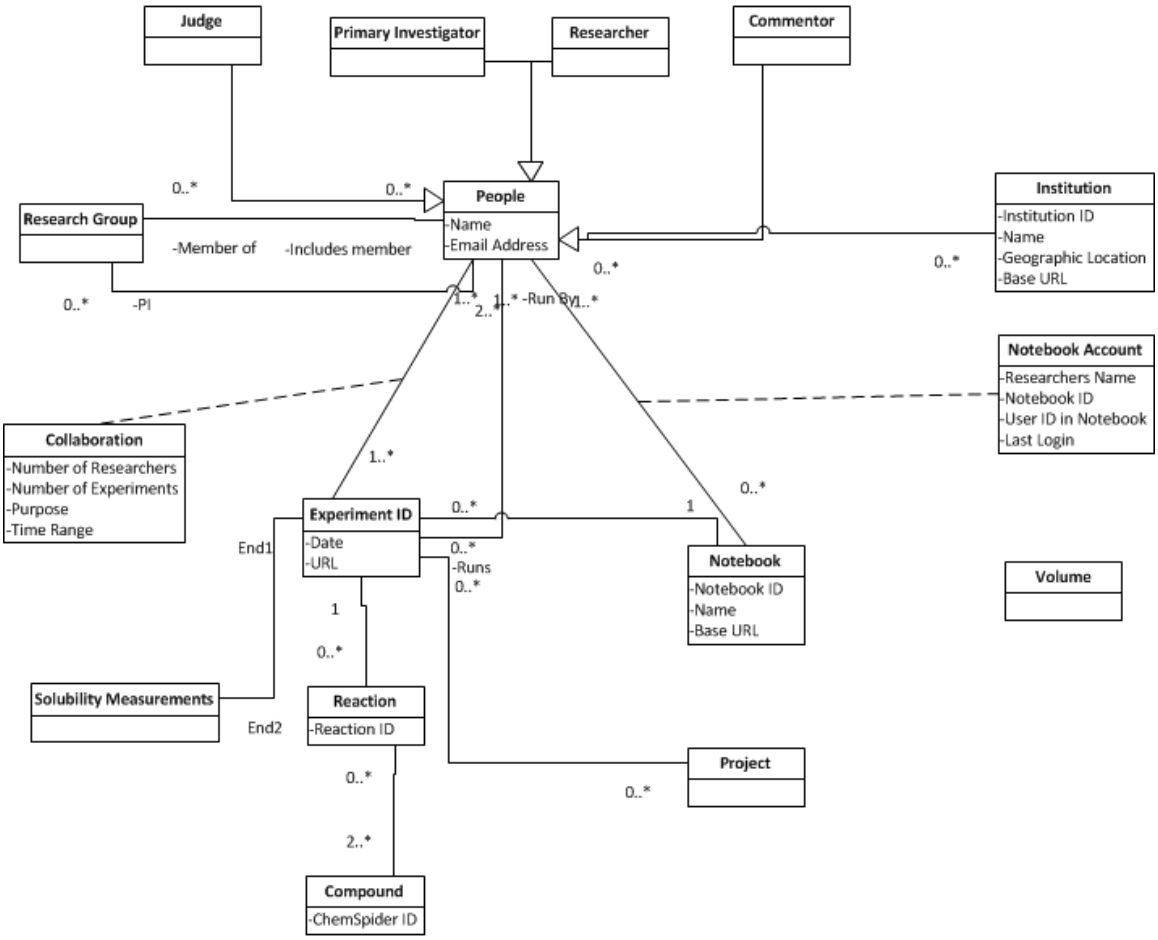


Figure 18: Inventory and model some of the core UsefulChem and Open Notebook Science data.

Creating a semantic network in the domain of Open Notebook Science as practiced for organic chemistry depends upon the available collections. Figure 18 attempts to inventory the core

entities and their attributes within this domain. The figure is represented using UML static structure notation. This model is specialized for the content of the Open Notebook Science Challenge.[23] Roles that individuals may hold include Judge, Primary Investigator, Researcher, and Commentator. These roles can be generalized as People who would have name and email address attributes that could be recorded. People may be affiliated with Institutions and Research Groups. They may maintain Notebooks and perform Experiments. Collaborations describe how People work together. Compounds may be used in Reactions that are performed as part of an Experiment. A series of Experiments may be run within the context of a Project.

The model in Figure 18 can inform decisions about which nodes should be instantiated from data. It also can be used to understand how the traversal of edges in the graphs corresponds to relationships and interactions within the domain.

The Social Molecule View

An experiment was performed to create a social molecule view. This view will enable exploration of the relationships between researchers and the molecules they have worked with. The relationships are established from ONS reaction records. We use the terminology of molecule and compound interchangeably. A log of this experiment was recorded on the Open Notebook Science Challenge site hosted by Wikispaces.[24] The same site is also used for the wet-lab experiment records for the ONS Challenge.

Two structured data sources were used for this experiment. Both of them are published as Google Spreadsheets. All of the reactants and products for the reactions recorded in the UsefulChem open notebook pages are recorded in a structured form in two spreadsheets. These

[23] http://onschallenge.wikispaces.com/
[24] http://onschallenge.wikispaces.com/DC-Exp-001

are the "ReactionAttempts"[25] spreadsheet and the "RXIDsReactionAttempts"[26] spreadsheet.

ReactionAttempts includes eight columns:

1. ReactionID – An identifier for the reaction.
2. Reference – A link to the document describing the reaction.
3. CompoundName
4. CSID – The ChemSpider compound identifier.
5. SMILES – A text representation of the compound's structure in the standard SMILES format.
6. Role – Indicates the role of the compound in the reaction, such as reactant or product.
7. Type – A generalization of the compound, such as aldehyde or amine.
8. SolventPredict

RXIDsReactionAttempts includes eleven columns:

1. ReactionID
2. Hyperlink
3. Precipitate
4. Product
5. Yield %
6. Researcher
7. Solvent

---

[25] https://spreadsheets.google.com/ccc?key=0Ak1R8T6wt4YQdG9NejNLcDNUMkVBVURGM01TR0NxdXc&hl=en

[26] https://spreadsheets.google.com/ccc?key=0Ak1R8T6wt4YQdGVENVFMWjdzaGd2REJTTnA4RG5vblE&hl=en#gid=0

8. Concentration of limiting reactant (M)
9. Notes
10. Reaction Type
11. Number of Reactants

The authoritative links for these data sources are available on the ONSbooks, Reaction Attempts page.[27] The Reaction Attempts Database is documented as:

> *"In order to enable the tracking of other types of reactions, the information in the CombiUgiResults sheet was reformatted into two other sheets: ReactionAttempts[11] (containing reagents and reactants) and RXIDsReactionAttempts[12] (containing reaction conditions and results, such as solvent, concentration of limiting reactant, appearance of a precipitate, yield, etc.). The two sheets are connected via the use of a common ReactionID. This format permits the representation of any type of reaction, with an unlimited number of reactants and products.[13]*
>
> *By definition, any Open Notebook Science project is a work in progress. The listing of a reaction in this database only means that the researcher attempted or is in the process of attempting it. Whatever the situation, a link to the laboratory notebook page is provided, where the most recent information is available. The philosophy used here is that partial information is always better than no information at all. Thus a researcher investigating the prior use a particular reactant in a Ugi reaction might find the report that a precipitate was obtained in methanol helpful for designing*

[27] http://onsbooks.wikispaces.com/Reaction+Attempts

> *their own reactions, even if the characterization of the precipitate is still pending. At the very least, knowing that a certain researcher has at least attempted a similar reaction is enough information for initiating a discussion, which may lead to valuable insights(Bradley, Mirza et al. 2010).”*

Live versions of the RXIDsReactionAttempts and ReactionAttempts Google Spreadsheets were downloaded as Microsoft Excel spreadsheets to the project DC-Exp-01 folder. The sheets from the two workbooks were combined into a single workbook. A VLOOKUP function was used to find the researcher names for each compound (CSID) in the data. A quick manual view of the results showed that some researcher names have multiple values in the same cell, separated with a forward slash character. Additional processing would be required to break these apart. The data structure creates a possibility to weight the edge by the number of ReactionIDs that establish the connection. Edges could also be directed with reactants on the left and product on the right. The lab can be extracted by looking at the prefix of the reaction ID.

The data was loaded into Gephi to create the graphs. Initial imports into Gephi did not complete fully. Although the Gephi Import CSV dialog did not report any errors, a manual sampling revealed missing records. The Context tab in Gephi reported only 738 nodes and 1024 edges although the edge worksheet includes 3941 records. It seems that Gephi will not load multiple edges between the same source and target. For example, Dustin Sprouse worked on molecule 7146 in at least reactions DSp35 and DSp36-1. Only the first edge was loaded into Gephi (DSp35). Due to this limitation is seems that edge weighting will be necessary to account for a researcher working with the same molecule in multiple experiments. Still clustering should be accurate in Gephi even if not all edge records are processed, since each molecule / researcher relationship will still be represented.

Average Degree	2.78
Network Interpretation	Undirected
Average Path Length	2.83
Number of shortest paths	510708
Graph Density	0.004
Modularity	0.465
Number of Communities	29
Weekly Connected Components	4

Table 4: Social Molecular Graph Statistics as reported by Gephi.

Table 4 reports some of the graph statistics that describe the overall structure of the resultant graph. The full overview of the graph's projection is shown in Figure 19.

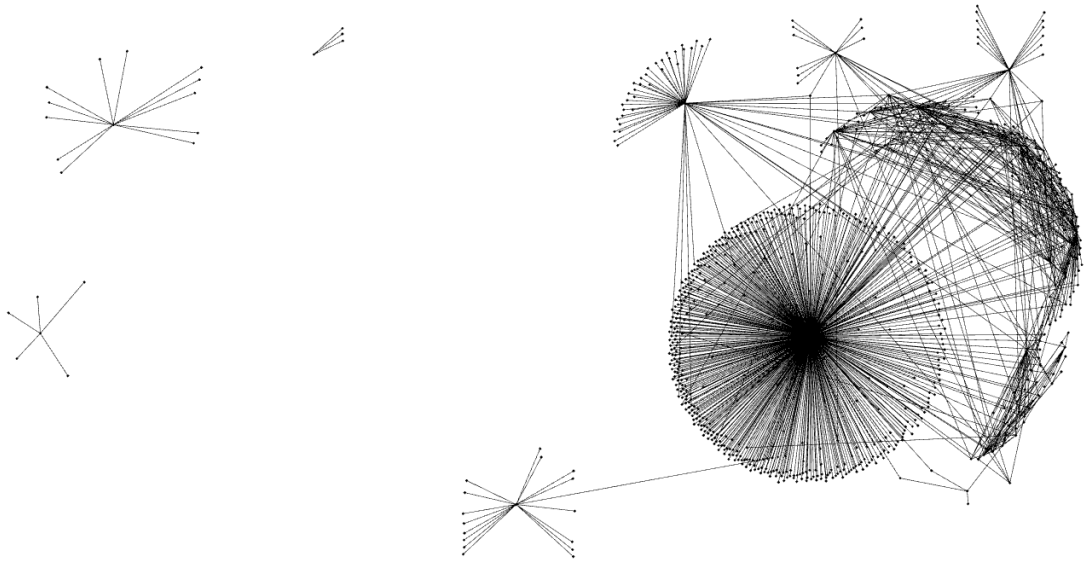


Figure 19: Overview Graph.

The overview shows a primary cluster of Ugi reactions centered on Khalid Mirza. There are also three disconnected clusters and five small, loosely connected clusters. The three disconnected

clusters are centered on a Khalid Mirza and Marshal Moritz collaboration, Dustin Sprouse, and Sebastian Petrik. These are shown in Figure 20, Figure 21, and Figure 22.

Disconnected Clusters

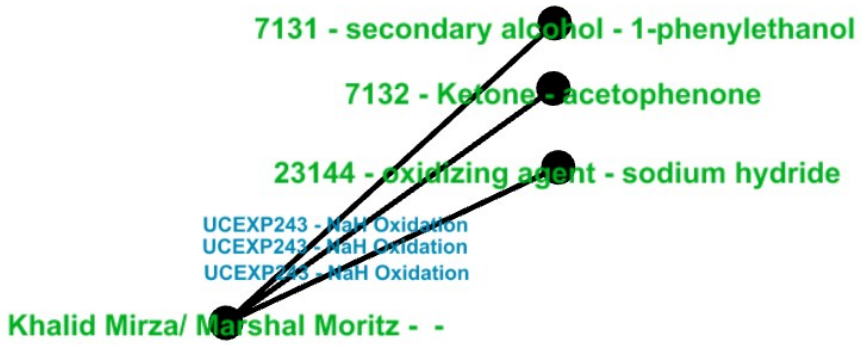


Figure 20: A disconnected cluster Khalid Mirza - Marshal Moritz cluster.

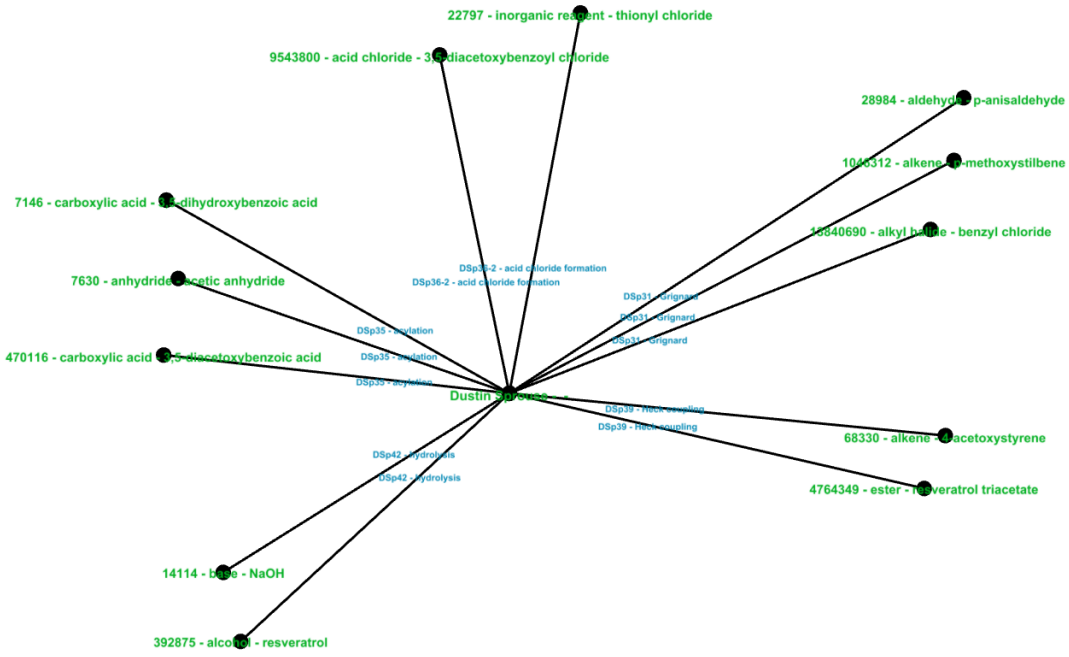


Figure 21: A disconnected Dustin Sprouse cluster.

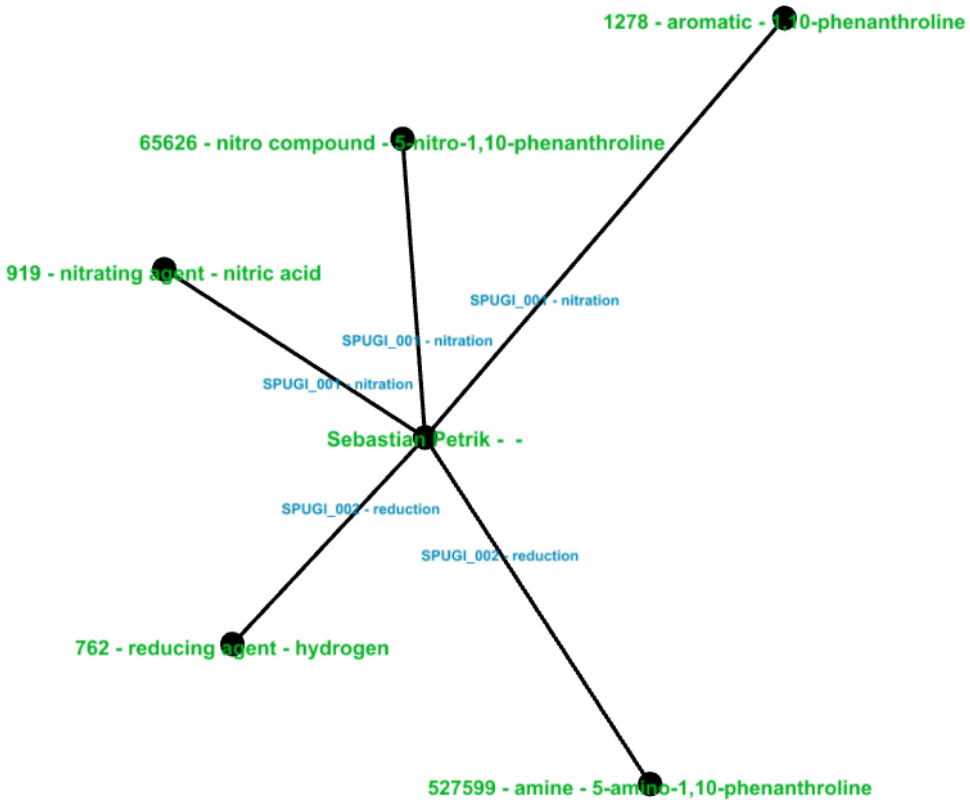


Figure 22: A Sebastian Petrik cluster.

Loosely Connected Clusters

The four small, loosely connected clusters are centered on David Bulger (Figure 23), a Khalid Mirza and Aneh collaboration (Figure 24), Marshall Moritz (Figure 25), a James Giammarco – Jessica Colditz / David Bulger – Khalid Mirza connections group (Figure 26), and Michael Wolfle (Figure 27).

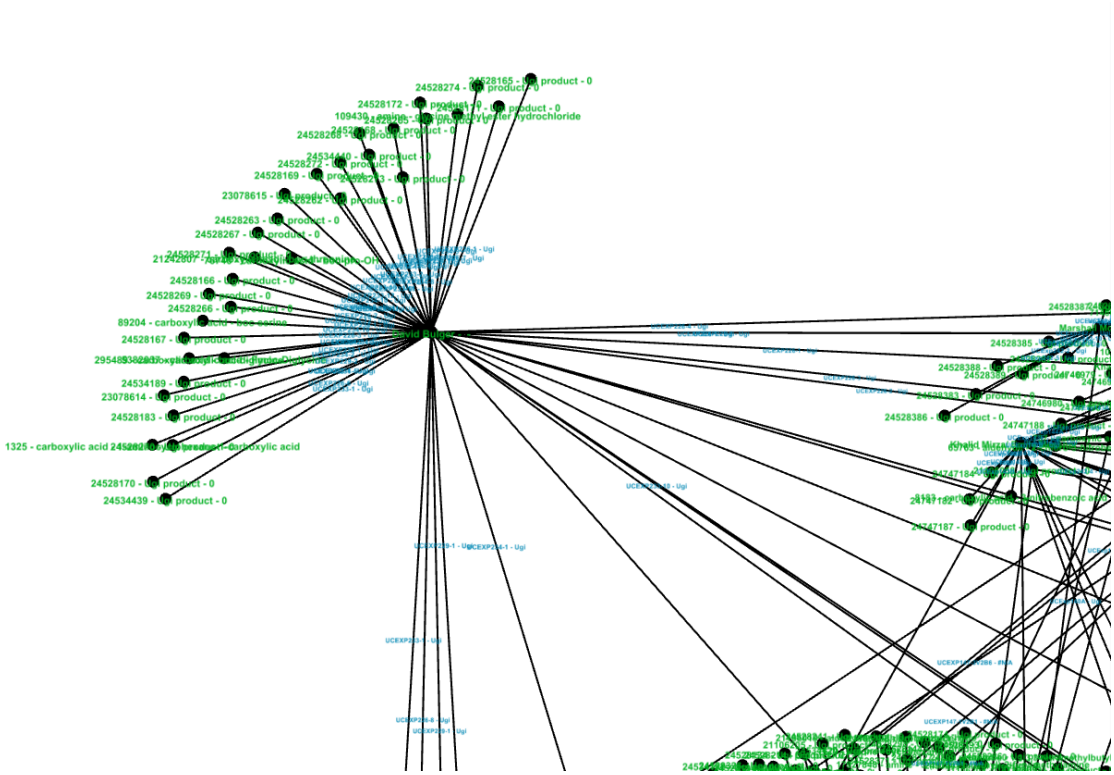


Figure 23: David Bulger cluster.

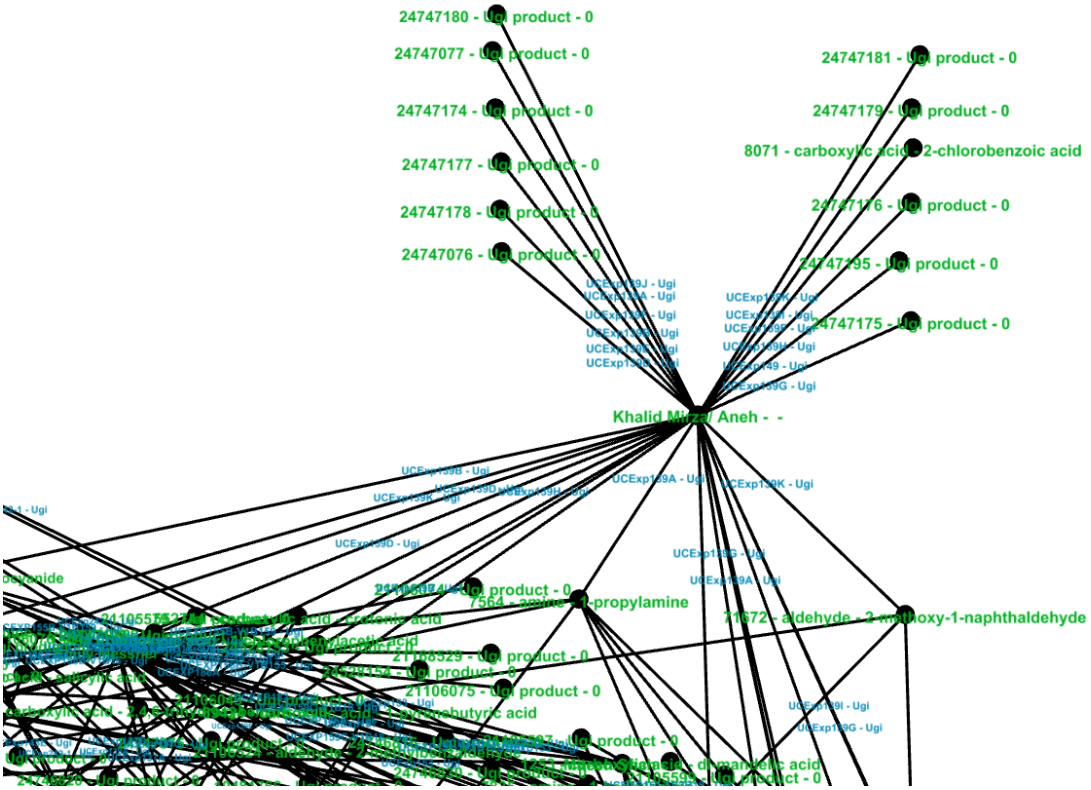


Figure 24: Khalid Mirza - Aneh cluster.

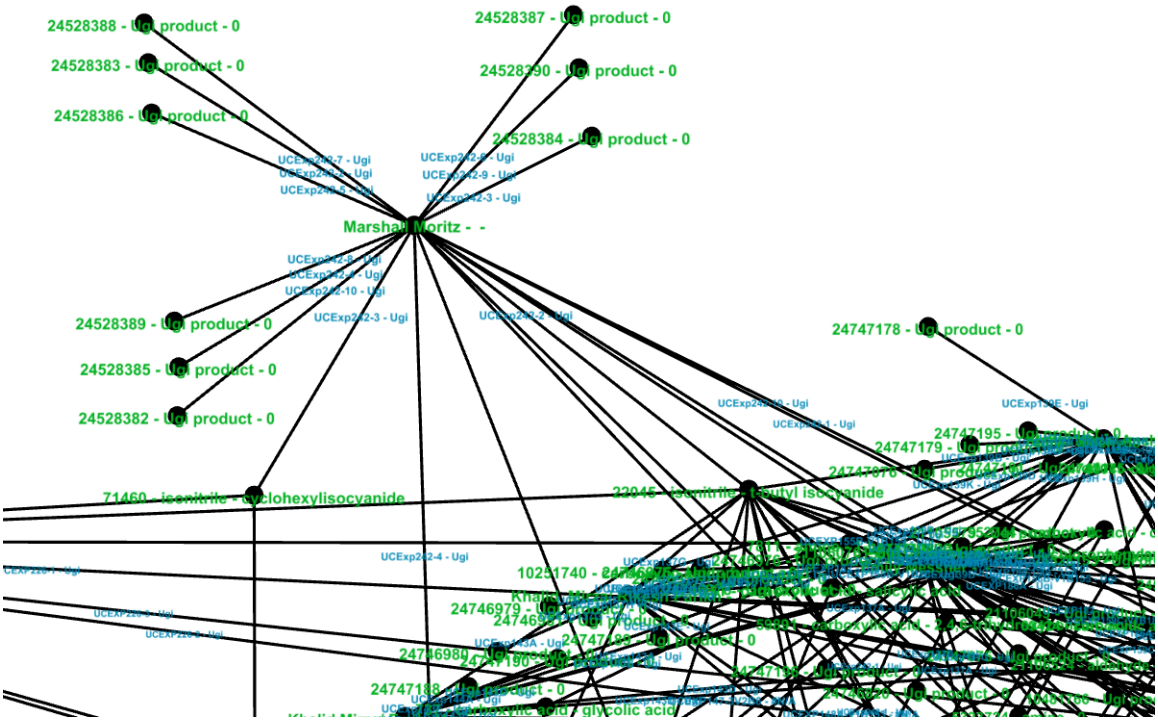


Figure 25: Marshall Moritz cluster.

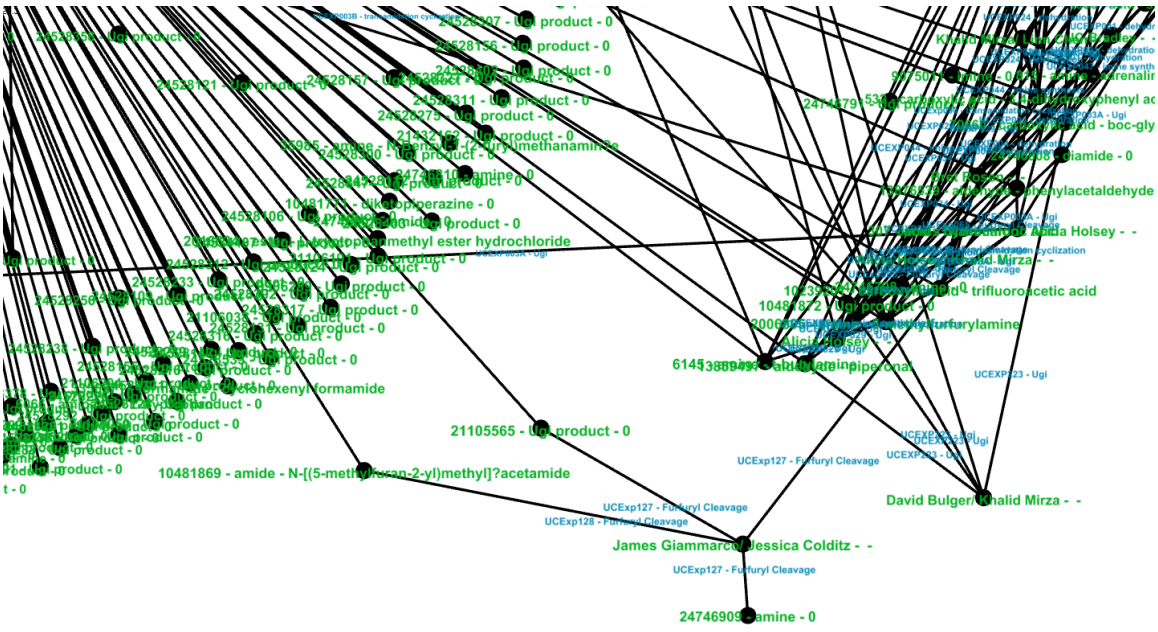


Figure 26: James Giammarco - Jessica Colditz and David Bulger - Khalid Mirza connections group.

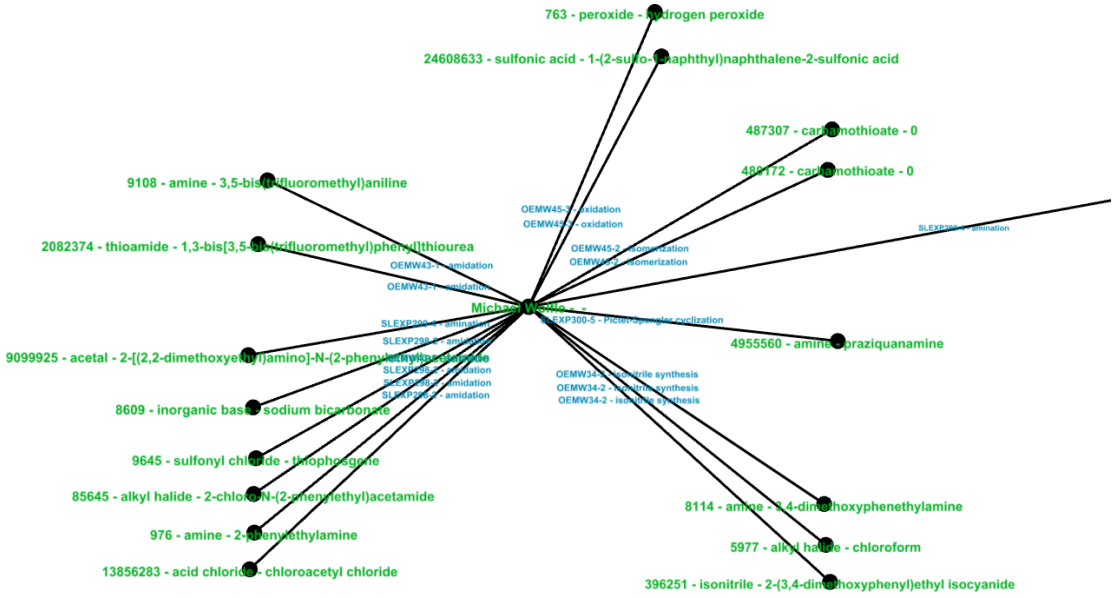


Figure 27: Michael Wolfle cluster.

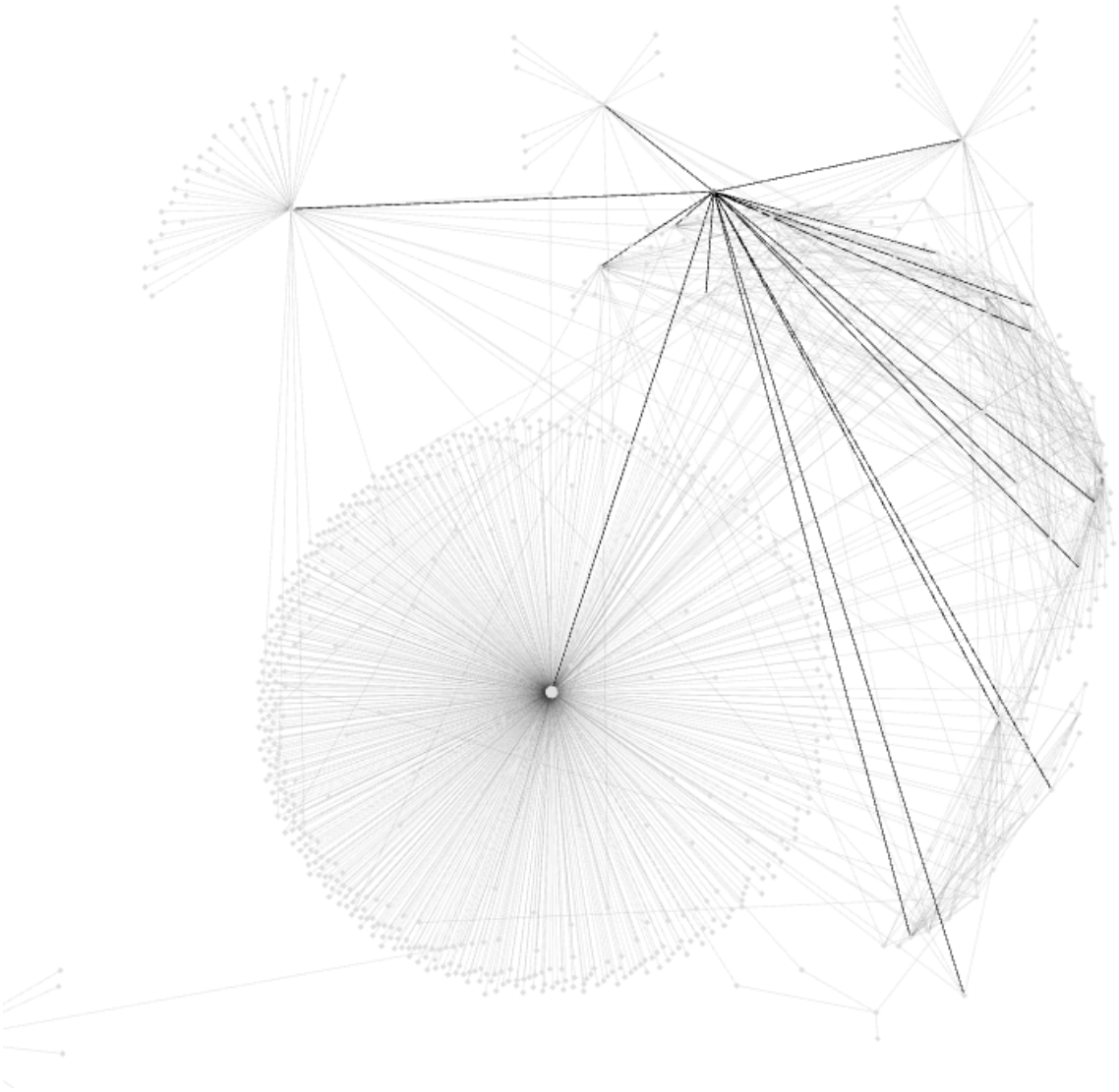


Figure 28: t-butyl isocyanide (CSID 22045) connections. The connections are highlighted in black with the fuller graph shown in lighter gray.

Gephi also provides a brushing feature that is useful for highlighting connectivity through dense areas of the layout and projection. Figure 28 shows the role of t-butyl isocyanide. It is common on multiple areas of work.

Observations

Figure 27 and the overview highlight that compound 80987[28] is the sole compound linking the Synaptic Leap[29] notebook with the UsefulChem notebook. That this is a valid link between the two notebooks is confirmed by checking the Reaction Attempts Explorer (Andrew Lang)[30] for aminoacetaldehyde dimethyl acetal. This link can also be viewed on the Reaction Attempt Advanced Search.[31]

The addition of the Synaptic Leap reaction data to the Reaction Attempts Database was reported as having been completed on May 2, 2010 by Jean-Claude Bradley in the post, "The Synaptic Leap Experiments to Reaction Attempts."[32] All of the imported Synaptic Leap reactions involved intermediates in the synthesis of praziquantel and were performed by Michael Wolfle under the direction of Matthew Todd at the University of Sydney. Praziquantel is a drug used to treat schistosomiasis, a disease caused by a parasite worm.

> *"Schistosomiasis is one of the most burdensome of the neglected diseases, with 200 million people infected and 400 million people at risk. Infection is widespread with a relatively low mortality rate, but a high morbidity rate, causing severe debilitating illness in millions of people. The disease is often associated with water resource development projects, such as dams and irrigation schemes, where the snail intermediate hosts of the parasite*

[28] http://www.chemspider.com/Chemical-Structure.80987.html
[29] http://www.thesynapticleap.org/
[30] http://onswebservices.wikispaces.com/reactions
[31] http://showme.physics.drexel.edu/onsc/reactionattempts/advancedsearch.php?compound=80987
[32] http://usefulchem.blogspot.com/2010/05/synaptic-leap-experiments-on-reaction.html

breed. The drug of choice for the treatment of schistosomiasis is praziquantel (PZQ). [Matt Todd][33]"

A month later, on June 1, 2010, it was reported that, "Andrew Lang noticed that there might be a quick synthetic route to praziquantel via an Ugi reaction."[34] Further investigation by Jean-Claude Bradley revealed that UsefulChem Experiment EXP258[35] documented an Ugi strategy for synthesizing praziquantel. Experiment 258 had been considered a failed experiment because it did not yield a precipitate. Matthew Todd separately identified a patent that had been published using this Ugi strategy.[36]

The Social Molecule View was developed after this key reaction had already been discovered. Analogously to the evaluation strategy used in the VAST Challenge, we can interpret this key reaction as a hidden ground truth in the data. To evaluate the potential utility of the Social Molecule View we can consider whether it provides information that would have increased the likelihood of earlier detection of this linkage. We can also consider the failure of the method to identify this link as a false negative.

The overview graph in Figure 19 invites and analysis of the bridges between the loosely connected clusters and the core cluster. The zoomed region shown in Figure 27 contextualizes the bridging connection with reference to the researchers involved, the laboratory notebook references, and the compounds. The semantics of the graph do not explicitly lay out for the reader that a more efficient technique for the synthesis of a treatment for schistosomiasis has been found. Rather, the structural characteristics of the graph may lead the reader to a path of

[33] http://www.thesynapticleap.org/?q=schisto/community
[34] http://usefulchem.blogspot.com/2010/06/use-of-ons-to-protect-open-research.html
[35] http://usefulchem.wikispaces.com/Exp258
[36] http://www.thesynapticleap.org/node/317

potentially fruitful investigation. The justification for pursuit of the path can be found in terms of the data elements that compose it. It remains that the reader must recognize the potential of the path. This form of use is consistent with the expectation that a given domain will emphasize different knowledge sources. By creating the visualization from the data, the researcher is most familiar with, his own data, and providing semantic links to data he is unfamiliar with, the method should be well suited to balancing the discovery of new data with the contextualization necessary to recognize the potential of the discovery.

Lessons Learned

The Open Notebook Science Study demonstrated that the method of using a heterogeneous associative network built from multiple dimensions of data could be applied outside of the 2008 VAST Challenge domain and in the Open Notebook Science domain. It also demonstrated that as with the 2008 VAST Challenge, visual exploration of this semantic network could provide a useful means for uncovering a hidden ground truth within the data. In this case, the structural shape of the graph invited exploration of the edge that was instrumental in the discovery of an Ugi reaction for the synthesis of an intermediate for the synthesis of the socially important compound praziquantel.

CHAPTER 6: PFIZER DRUG DISCOVERY PROJECTS STUDY

Project Description

The drug discovery process involves many steps. Although the sciences evolve organically and iteratively, it can be useful to use a linear, funnel model to describe the drug discovery process. Although generalizations are made which may not be completely fulfilled in all instances, the funnel model provides a useful framework and vocabulary for discussion. Inputs and project inception activities constitute the left side of the linear funnel and outputs such as the manufacture and sale of commercial drugs constitute the right side. Project inception activities include exploratory research, disease targeting, and research strategy formulation. Tasks in this area include figuring out which questions to ask. The identification of starting points is a pervasive and critical task supporting inception activities.

Historical data from successful research projects at Pfizer Research and Development were collected and organized for study. The primary data object is a compound that was synthesized for experimentation during the course of a defined drug discovery project. To protect confidential information, the molecular structure for each compound was not provided in the dataset. Instead, various meta-data were provided. The most central of these were similarity scores indicating the structural similarity between two different compounds. Drexel Identifiers were defined as unique, custom identifiers for compounds. Drexel Identifiers (DXID) include a constant three-character prefix (“DX-“) followed by a ten digit number. The ten-digit number is unique for the compound. The Pfizer team members maintain a mapping from the DXID to the identifiers in Pfizer R&D systems that can be used to open the full record for the compound.

We approached the data by graphing the compound records as nodes and defining edges indicating the similarity between two compounds. This provided a mechanism for representing

the data graphically. From there, the graphs and visualizations were refined to make them more useful for answering questions about the project. It was hypothesized that exploration of the data in this way could lead to statements about the role of the compounds during the course of the R&D project. To examine this hypothesis, key compounds were identified in the project and then compared to their position in the resulting graphs. The definition of key compounds is subjective. Researchers who worked on the project could reflect on the history and identify compounds that were remembered as having played critical roles in the project evolution. An objective measure of key compounds was sought that could approximate this subjective definition. The objective measure provided consistency for evaluation of techniques for analyzing the graphs. The team of Pfizer and Drexel researchers collaborating on this study felt that enumerating the compounds that were explicitly listed within the PowerPoint slides written by the project researchers would be a suitable objective measure.

Five Pfizer R&D drug discovery projects were assessed for study. These were assigned identifiers of A, B, C, E, and Z. Due to differences in how the projects were run and the information artifacts that they produced some projects have nuances in their representation that differ from others. Therefore, between-project analyses must be handled with care. Additionally, within-project analyses are understood best by interpretations that account for the project-specific nuances. A Microsoft SharePoint site was used as a central repository for the project files. This site provided version control and provenance information for the file. The file names from this repository as used as identifiers for subsequent references.

Project B Key Compounds

A list of DXIDs found in the PowerPoint slides of project presentation was defined. This is not expected to be an exhaustive list of all relevant compounds. These will be helpful for

understanding a bit of the story; however, they will not include all compounds that may be of interest. This list serves as an objective measure that approximates the subjective list of key compounds from the R&D project.

DX-1955354783
DX-7964907477
DX-9408546408
DX-0744597840
DX-6962797697
DX-0861470606
DX-1348271437
DX-6796085261
DX-5573569428
DX-0380309731
DX-4154668781
DX-9512580217

Table 5: Project B nodes that occur in a PowerPoint presentation given by the project team.

Methodology

This section describes the process of translating from Pfizer database data to the DXID data collection used for these analyses. All compounds registered to Project B on were selected based on the assignment made at the time the compounds were registered. The Pfizer IDs were assigned an arbitrary unique DXID that could be used to reverse lookup the Pfizer ID. The date the compound was first registered to the project was also collected. This process produced a node file with the DXID and registration date. The following fields are in the nodes file:

- REG_DATE – The date the compound was registered in the Pfizer database, the Research Information Factory (RIF) 2.
- DXID – The Drexel identifier.
- NumberOfClosest - This field is not valid. It is an artifact of the process that was subsequently not used for the analysis.

Each compound was compared to all compounds registered on a prior date, and the similarity (Tanimoto based on ICFP_6 fingerprints) calculated. All similarities of greater than 0.65 were retained as edge connections. This process produced an edge file a "from" field identifying a DXID having a similarity measure greater than the threshold and having been created prior to the DXID in the "to" field. The following fields are in the edges file:

- FROM
- TO
- Similarity

The Project B collection consisted of 4,445 nodes and 29,884 edges. A number of tools were used to explore the resultant graphs. Maple was used to process data however, the graphs were found to be too large for the interactive graph visualization capabilities of Maple. CiteSpace and Gephi were found to produce good layouts and have good interactive performance. Spotfire was also used. Graph layouts performed in CiteSpace or with the Gephi implementation of the OpenOrd (Martin, Brown et al. 2011) algorithm were loaded into Spotfire. Spotfire provided a mechanism to enable interactive brushing of nodes and display of Pfizer structure data in coordinated views.

Expert Feedback

Jared Milbank of the Pfizer team and I met with a Pfizer Researcher who worked on Project B. We met in the Researcher's office. This Researcher has been involved with Project B for the duration of the project. He is a senior researcher with a high degree of experience and expertise.

A laptop was brought to the office to display the visualizations. Spotfire[37] was used for the displays. At points in the interview, screenshots were taken. A transcript of my notes from the interview is provided in the Appendix. The screenshots and notes have been integrated in this section to describe the responses. Using Spotfire on Jared's laptop had the advantage that the DXIDs could be mapped to the Pfizer identifiers in real-time. The DXIDs did not have any meaning for the researcher, although he was very familiar with the internal Pfizer identifiers for the compounds. References to DXIDs within the notes are recorded using the last three digits of the ten-digit identifier. This made it easier to keep up with the discussion while taking notes. The Researcher also mentioned that this is a typical way to refer to long identifiers within the organization.

Jared and I also met directly after the session to discuss the responses. This helped to ensure that we had interpreted the feedback accurately. We also met with the Pfizer-Drexel Collaboration Team later that day.

The session was organized into three sections. The purpose of the first section was to provide overview materials and to get initial reactions from the Researcher. I was introduced by Jared. Jared and the researcher were already acquainted and had discussions on this project in the past. I described this study in the context of the Pfizer-Drexel Collaboration. We then described the Collaboration's objectives and high-level strategy. After this, we asked the Researcher if he had any questions. We then asked the Researcher for his opinions on the approach.

The purpose of the second section of the session was to gather feedback on the potential use of the graph-theoretic metrics. We presented the Researcher with some of the visual

[37] http://spotfire.tibco.com/

representations in coordination with nodes that had been scored according to indegree, outdegree, and betweenness.

The third section was similar to the first, and modeled loosely on a pre-test / post-test assessment style. After the Researcher explored the visuals and scored nodes of Project B data during the second section, we then asked more general overview questions. We asked the Researcher for his opinion on the general approach. We also asked if he thought it had potential utility, and if so what kind.

The questions were designed to elicit feedback on the utility of the methods. The format was open-ended to provide as much opportunity as possible for the Researcher to offer feedback without prompting from us. The questions that were asked were guided to detect three cases of utility:

- No Utility: Failure of the system and methods.
- Incremental Utility: New perspective on existing ideas.
- Strategic Surprise: An “ah-ha” moment or disbelief.

The session would have detected the case of “No Utility” if any of the following had been perceived: (1) existing methods that already do the same thing, (2) explicitly being told that there is no utility or, (3) low comprehension of the presented visualizations. The combination of open-ended questions to elicit feedback and explicit questions designed to detect these three cases were used though out all three sections of the session.

The criterion for detecting “Strategic Surprise” case included perception of an “ah-ha” moment and/or shock and disbelief as what was being seen. Responses where the Researcher saw a new

connection or pattern that was so significant that it began to dominate the discussion would be indicators of the “Strategic Surprise” case.

Detection of the “Incremental Utility” case included perception of enthusiasm for the system and methods and unsolicited ideas for use. The Researcher offering suggestions for how these systems and methods might be used would be indicative of the “Incremental Utility” case.

Section 1

After introductions, we provided a high-level summary of the project. We broadly described our approach of modeling the compounds and their similarity with a graph. Before providing details, we asked questions to detect the “No Utility” case. We asked if the Researcher was aware of existing tools or methods that do the same thing. He was not aware of any. We asked the Researcher if he thought the strategy had merit and he agreed that it sounds like a reasonable approach. We also prefaced these questions with more open-ended questions such as “What do you think?” to provide an opportunity for the Researcher to offer feedback. The Researcher did not offer any information that gave support to the “No Utility” case.

We explained that the input data for the visualizations consisted of compound identifiers and structure similarity scores. The structures for the compounds were not used. This had been done to protect Pfizer intellectual property. A graph was created represent the compounds. Compounds were represented as nodes in the graph and these were represented as circles on projections of the graph. Edges were instantiated in the graph from nodes that were registered later in the project back to nodes that were registered earlier in the project if the similarity score between the two nodes was above a given threshold.

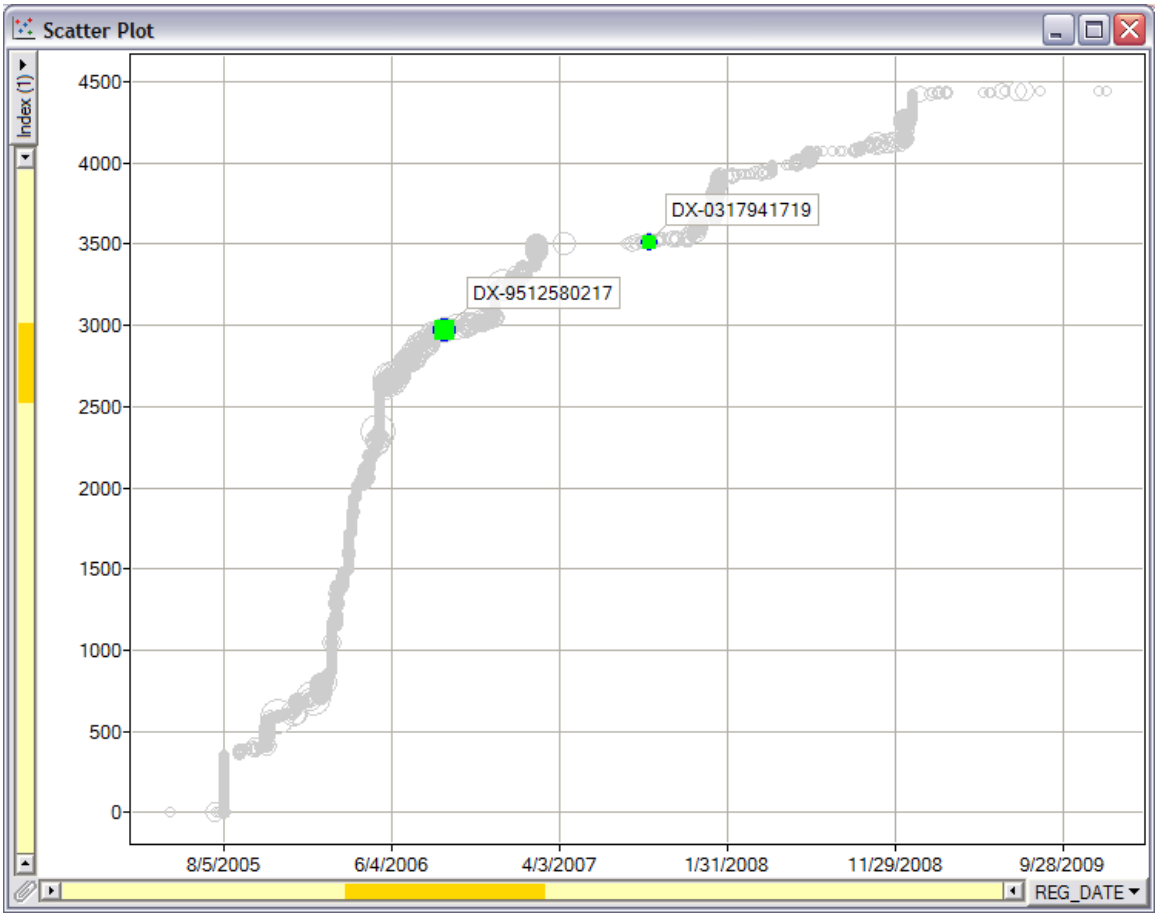


Figure 29: Timeline view of compound identifiers by the date that they were registered in the compound database.

The first visualization shown was the timeline view. A screenshot is shown in Figure 29. The two candidates produced by this project are highlighted in green. The Researcher recognized that the project had been run almost as two separate projects. The first collection of data supporting the initial compound is shown before the temporal gap in the figure. The second collection was created to support the development of a backup candidate.

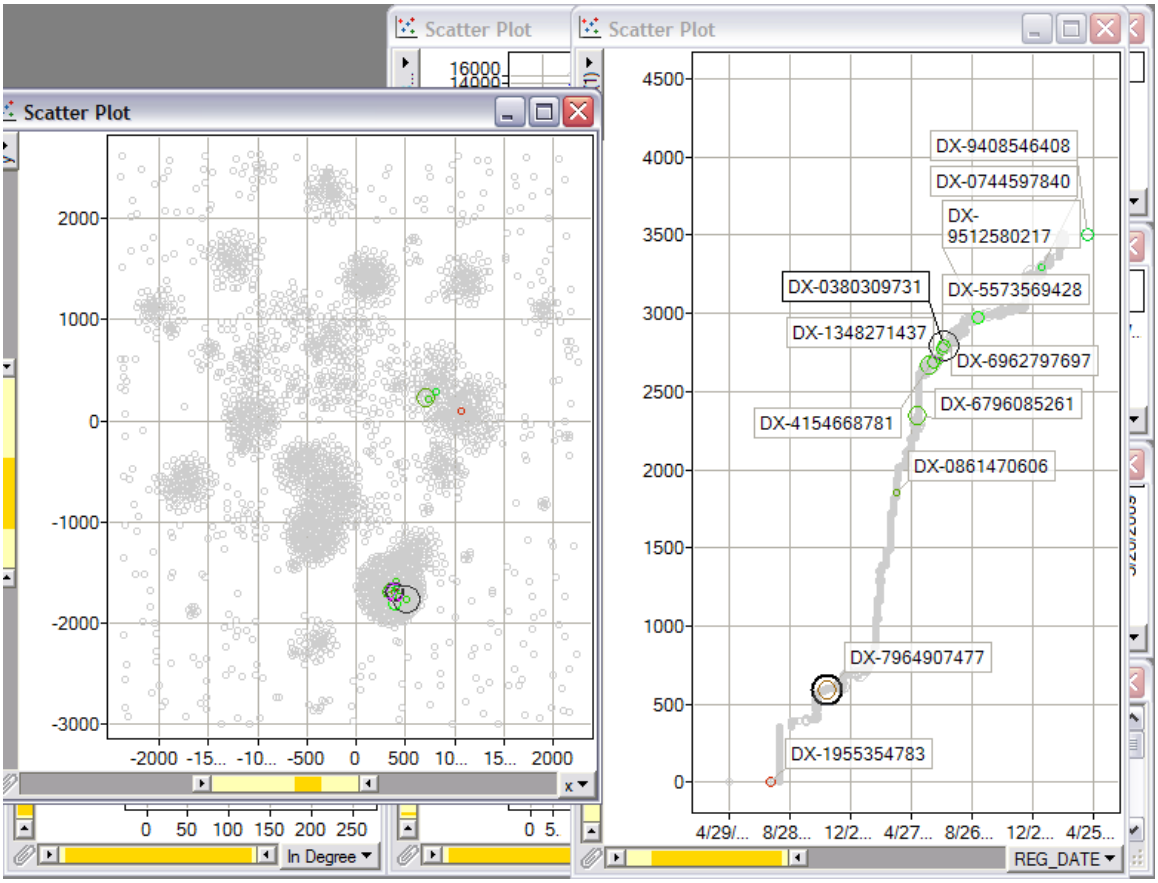


Figure 30: Coordinated views of clusters and the timeline.

A screenshot of the second visualization presented is shown in Figure 30. In this visualization, the frame on the left displays a projection of the full graph. Nodes are drawn as circles. Edges are not represented. The frame on the right shows the timeline. The timeline has been zoomed to show the first phase of the project done in support of the first candidate compound. Nodes for the compounds that appear on PowerPoint slides from the project are colored in green in both views. They are also labeled in the timeline view. In the left frame, relative distance between two nodes indicates structural similarity. Circles that overlap or are near each other represent structurally similar compounds. Circles that are far apart represent dissimilar compounds.

The Researcher indicated that there was value to having a full overview of all of the compounds in a project as shown in Figure 30. He mentioned that it is common to keep a few people constant and on a project for its duration. It is also common to bring a few people on and off during the course of the project. The constant people provide continuity while those entering new bring fresh ideas. He mentioned that a full overview is not currently available to people moving on and off a project. He felt that they might benefit from such an overview. He also felt that it would be useful to correlate the clusters with biological data. For example, it would be useful to answer the question, "Does one cluster give you the desired biological properties more often than not?" He also mentioned that big clusters of connectivity made sense. They represented the libraries that were run for the project. Spotfire provided interactivity with the visualizations. Brushing, zooming, and scaling operations were available. The Researcher spent some time interacting with the tool. As he interacted, he exhibited a process of formulating expectations and then using the data to verify those expectations. This indicated that he understood what he was seeing. During his interaction, he remarked that it could make you ask, "What did you get out of these clusters – Should we have spent so much time there?"

Section 2

After the researcher had explored the visualizations shown in Figure 29 and Figure 30, we asked open-ended questions again. We also asked if he had any questions in general. He was quite comfortable with the approach so far and had provided feedback indicating that the visualizations were comprehensible. We then explained that the next visualizations would show the compounds scored against measures derived from the underlying graph structure.

Indegree

The indegree for a node is the count of the number of edges that terminate at the node. In establishing the directed edges for this graph, an edge was instantiated from a node to nodes for compounds that were made earlier. This was only done if their similarity was within the threshold. Therefore, the indegree can be interpreted as the number of similar compounds that were registered later.

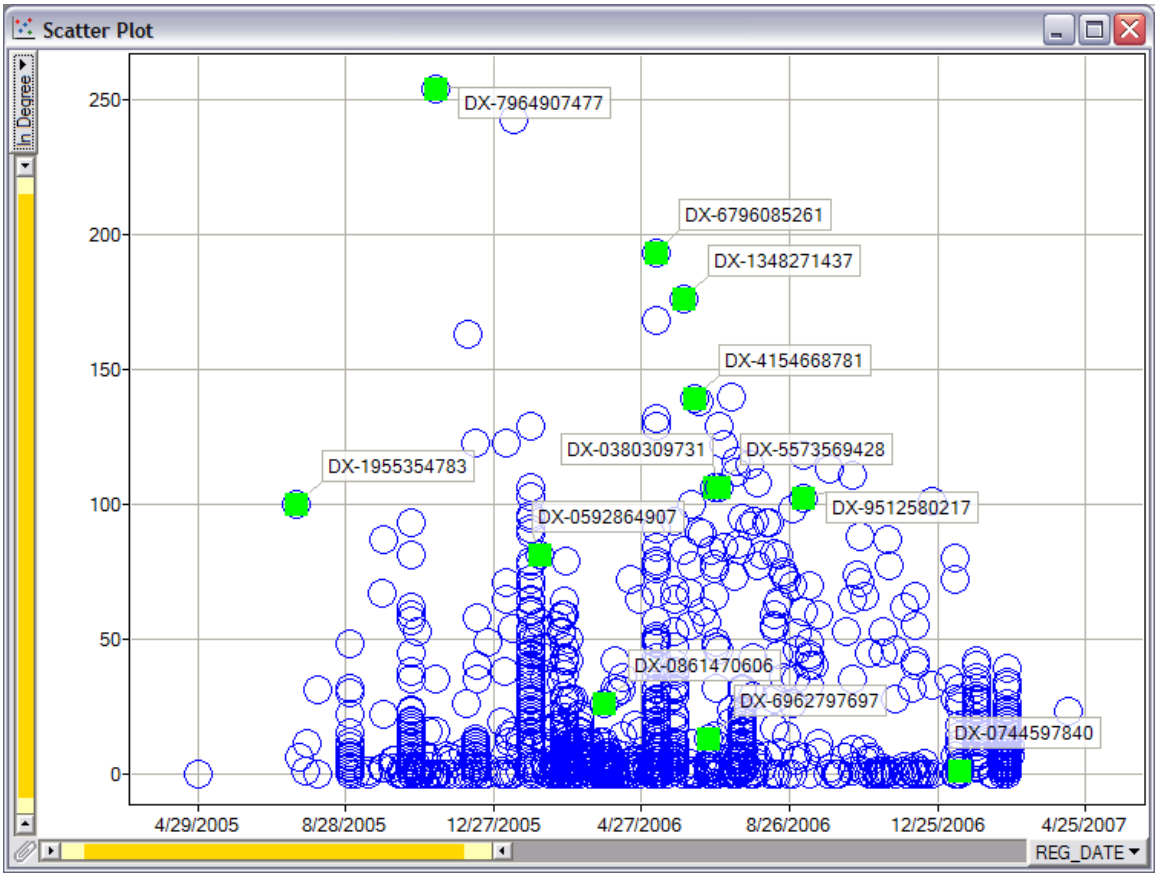


Figure 31: Screenshot of indegree view.

Figure 31 shows the date that a compound was registered with Pfizer's internal system on the x-axis and the indegree of the node in the graph structure on the y-axis. Circles representing nodes for compounds that appeared in the PowerPoint slides for the project are colored green and labeled.

It was observed that a fair number of compounds that were important enough to be mentioned in the project summary slides also had a high in-degree. The indegree was found useful for answering the question, how many compounds did we make like it? It was observed that the left-most green compound, “DX-1955354783” was the initial lead. As an initial lead many subsequent compounds were made that were structurally similar to it. The Researcher remarked that this view would be related to the use of library chemistry. “DX-7964907477” was from the first library. It has the highest indegree score. This visualization prompted the question; why was the node with the second highest indegree not mentioned in the slides?

Outdegree

The outdegree for a node is the count of the number of edges originating from the node. In establishing the directed edges for this graph, an edge was instantiated from a node to nodes for compounds that were made earlier. This was only done if their similarity was within the threshold. Therefore, the outdegree can be interpreted as the number of similar compounds that were registered earlier.

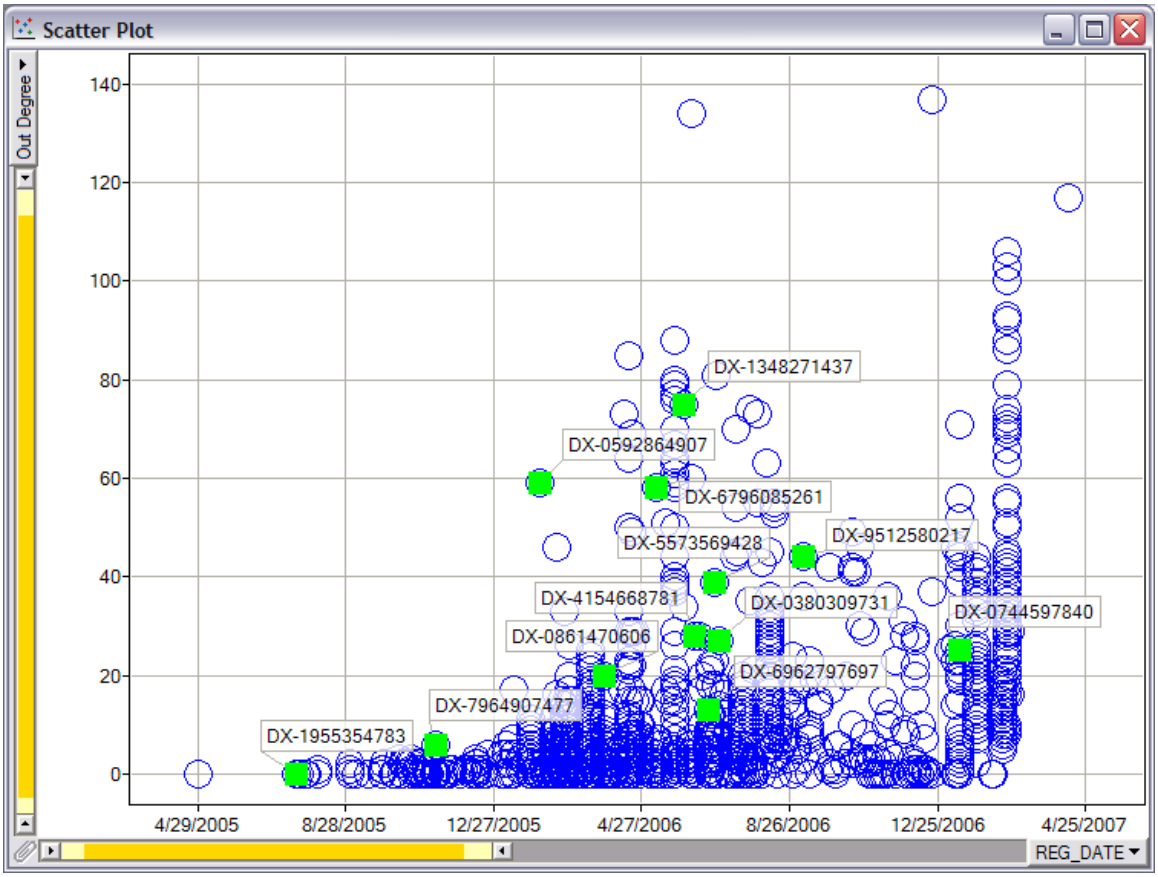


Figure 32: Screenshot of outdegree view.

Figure 32 is similar to Figure 31 however; the y-axis shows the outdegree rather than the indegree. Having a high outdegree indicates that many compounds were made that were similar to the one with the high outdegree. The Researcher used brushing to explore compounds with high outdegree and recognized some as having been used to answer very specific questions. He remarked that this visualization might be useful in a design meeting. He indicated that it could be useful in preventing the synthesis of a new compound if it is very similar to a large number of compounds that have already been made for the project. In such cases, the compound should be trying to answer a very specific question.

Betweenness

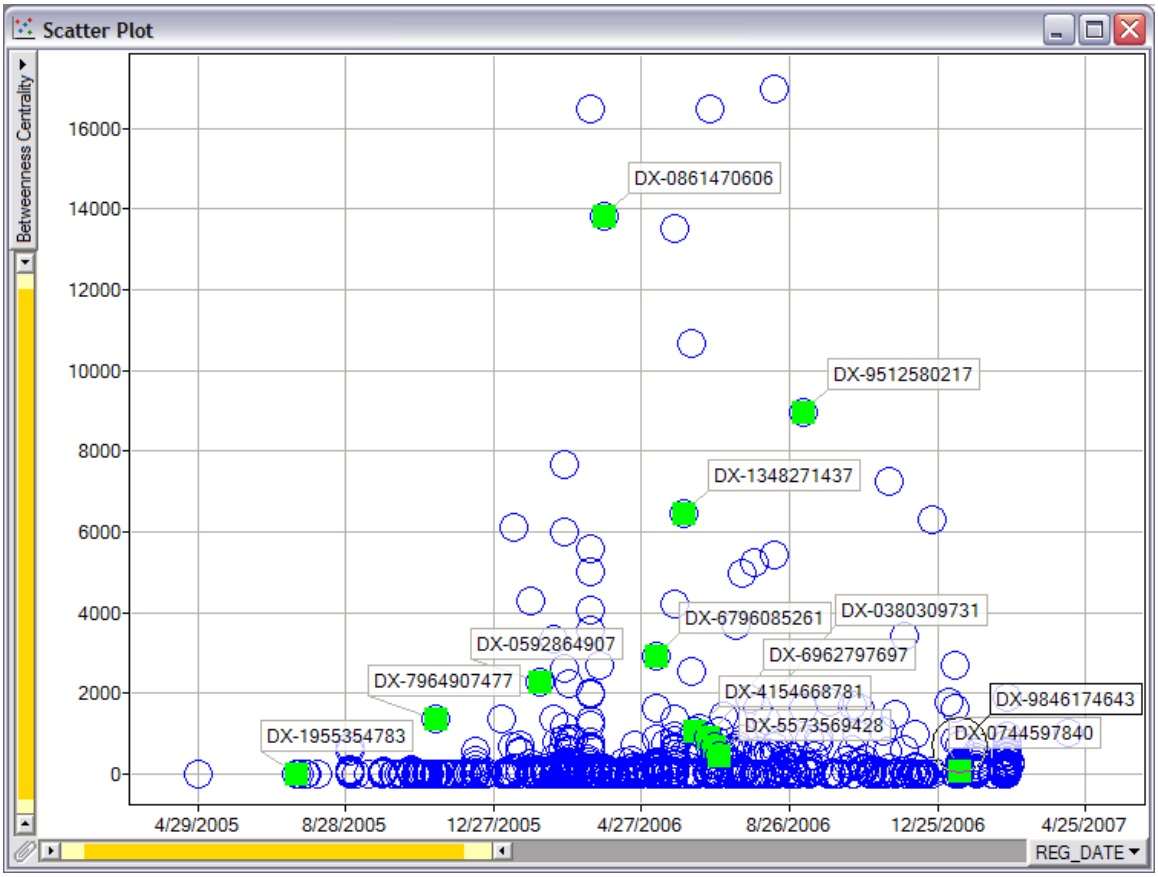


Figure 33: Screenshot of betweenness view.

Similarity to the previous two views, Figure 33 shows compounds scored by the betweenness metric. “DX-0861470606” stands out in this view as having a high betweenness score and being significant enough to include in the project summary slides. The betweenness scores seemed to generate less interest than the indegree and outdegree scores. It was thought that betweenness might be used to identify singletons versus library compounds but the utility of this approach seemed tentative.

Section 3

We reviewed the compounds identified in using the indegree, outdegree and betweenness metrics. We again discussed the general merits of the approach and asked for open-ended

feedback. The Researcher felt that the methods would be particularly useful for problematic projects. He felt that it would also be useful for someone who was joining an already running project. It was pointed out that these methods would only work when the drug discovery team was able to recognize key compounds. He felt that another use might be to highlight interesting nodes that have not previously been followed up on.

Lessons Learned

The Pfizer Drug Discovery Projects Study demonstrated that the method of using an associative network built from historical project data could be used to provide insight into the drug discovery process. This study extended beyond the visual exploration of the data that was performed in the prior studies. In this study, quantitative properties of the graph were used to identify records that merit investigation for researchers. Expert feedback indicates that indegree and outdegree hold potential utility for this purpose.

CHAPTER 7: CONCLUSIONS AND DISCUSSION

Trends in the Literature

Considering the history of scientific communications and indexing strategies that were discussed in "CHAPTER 2: Literature Review," we can see two parallel trends emerging. First, we see that technological advancements can lead to changes in both the kind of artifacts produced and the volume of artifacts produced. A chapter in the 2011 book, "Collaborative Computational Technologies for Biomedical Research" opens:

> *"Technology has a profound effect on how scientists can communicate with each other. This affects how quickly science can progress and what kinds of collaboration are possible (Bradley, Lang et al. 2011, p.426)."*

A result of this trend is that issues of information overload become acute and models of scientific communication may need to be revised. The second trend is a response to the first. We see that the introduction of new kinds of artifacts and increases in volume lead to advancements in the methods used for indexing. In 1949, the Army Medical Library had recognized the need to improve indexing methods in response to the increasing volume of medical literature:

> *"One of the most serious problems confronting science at the present time is the difficulty in keeping abreast of all the research that is being done and in bringing the published results into some workable order. If the results of research are buried or lost for some reason or other, the research, and the money spent on it, is entirely wasted. To prevent such a loss we need adequate guides to the vast amount of scientific literature and must make*

> *intelligent and effective use of them. … It is becoming increasingly difficult for our indexes and abstract journals to keep up with the growing number of medical publications and with articles of medical importance in other scientific journals. … The aspect of the problem which is our immediate concern today and which is particularly important to the Army Medical Library is that of the role of indexes in meeting the needs of the present and of the future (Larkey 1949).”*

An interaction effect between these two trends may create a cycle. Advancements in methods used for indexing may make more materials discoverable and reusable. This in turn may lead to technological advancements that then again push the limits of existing indexing methods. It appears that the technology advancement iteration is ongoing. Recent advancements include cloud computing, eScience, Data Driven Science, Open Notebook Science, Cyberinfrastructure, and Open Data initiatives. An emphasis in this iteration of the technological advancement trend is data. Together these advancements are pushing the limits of existing indexing methods. The methods described in the preliminary study and three experimental studies here provide candidate solutions to this pressing indexing problem.

Lessons Learned

The preliminary study of the 2008 VAST Challenge demonstrated that heterogeneous semantic networks created from multiple data sources could be useful for leading to new insights about a collection and for finding a hidden storyline. The limitation of this study was that it was performed on only one artificial problem. Additionally, it required substantial manual effort to encode the data making it difficult to apply to new or highly dynamic collections.

The Influenza Protein Sequence Study demonstrated that the graph data structure could be applied in a real-world scientific domain. It also demonstrated that the method of using the graph structure could be scaled from the small collections studied in the 2008 VAST Challenge to the real-world collection of protein sequence data. Limitations on interactive exploration of large graphs with current tools were identified and overcome with the construction of a new system. This system was then used for an analysis of the graph. Together these provided a new overview that provided novel insight on the dynamics of the collection. They also provided a new means for identifying interesting members of the collection that would be more difficult to identify using the form and filter methods of existing web form-based systems.

The Open Notebook Science Study further demonstrated that the method of using a heterogeneous associative network built from multiple dimensions of data could be applied outside of the 2008 VAST Challenge domain. This study demonstrated that as with the 2008 VAST Challenge, visual exploration of this semantic network could provide a useful means for uncovering a hidden ground truth within the data.

The Pfizer Drug Discovery Projects Study demonstrated that the method of using an associative network built from historical project data could be used to provide insight into the drug discovery process. This study extended beyond the visual exploration of the data that was performed in the prior studies. In this study, quantitative properties of the graph were used to identify records that merit investigation for researchers. Expert feedback indicates that indegree and outdegree hold potential utility for this purpose.

These methods therefore demonstrate a feasible approach to indexing that can address the volume and diversity of data being produced by recent technological advancements in data production. The heterogeneous semantic network has the characteristic of supporting

interactive visual projection for exploratory analysis. It also enables quantitative analysis. This combination has been shown to be promising for assisting users with the identification of key records and relationships.

Future Work

Round-Trip Engineering

A key enabler for discovery algorithms that process integrated collections of literature and data is the instantiation of links between entities referenced in the literature and records for those entities in structured databases. In the case of Open Notebook Science, explicit reference to ChemSpider identifiers from within the wiki text of a notebook page enables parsers to directly instantiate links in a graph data structure. This can be compared to the process of instantiating links between journal articles in PDF format and ChemSpider records for referenced compounds. Given PDF format, the process requires additional steps, each of which introduces uncertainty to the resulting graph structure. First, the text must be extracted from the PDF. Next, the text must be processed using NLP algorithms such as those implemented in the OSCAR API.[38] Finally, the results of the processing must be used as input to a web service search that ultimately identifies the relevant ChemSpider identifiers.

In "Model-Oriented Scientific Research Reports," the addition of structures created using scientific communication modeling techniques is proposed as a way to overcome limitations in analyzing narrative descriptions (Allen 2011). Under this proposal, inefficiencies in natural language processing of narrative could be overcome by fitting the report into well-defined and formally modeled structures. A challenge in implementing this proposal might be getting authors to produce model compliant reports. Although narrative descriptions are difficult for

[38] https://bitbucket.org/wwmm/oscar4/wiki/Home

computer to process, they are comparatively easy for authors to produce. Additionally, a robust market of word processing tools exists to support the creation of narrative in reports. The proposal notes that, “User tools could also be developed for authoring model-oriented research reports and for browsing the library (Allen 2011).”

The approach of adding structure to scientific research reports, and building user tools for authoring such reports, can be extended to consider the full work process. For example, when a Chemist designs a new reaction to be performed the reaction would normally be entered into the laboratory notebook. At that design time, the Chemist is concerned with the question, “Has anyone performed this reaction before?” The current state of the art requires that the Chemist switch to another system, such as Reaxys,[39] and submit a query. Submitting the query requires entering the definition of the reaction into the query form using the format required by the search system. Once satisfied, the Chemist then moves on to input the reaction into the laboratory notebook as the goal of the experiment. A tool that could integrate search with notebook data entry could combine these two steps. With this alternative, the Chemist could enter the reaction once, into the notebook. If the notebook includes the mechanism to perform the search, the Chemist would be motivated to enter the notebook entry in a structured way, to accommodate the search tool. This would have two benefits: (1) round-trip engineering would be achieved such that the structure most beneficial to the search is the same as the structure the Chemist is motivated to enter and (2) the process of answering the question “Has anyone performed this reaction before?” could be captured. This would both reduce the effort needed by the Chemist and capture the artifacts of the scientific inquiry more completely.

[39] https://www.reaxys.com

Currently, the Open Notebook Science wiki pages at least use ChemSpider Identifiers as annotations when molecules are referred to in the text. The wiki tool alone does not provide automation for the lookup of the appropriate ChemSpider Identifier, or for the inclusion of the identifier into the narrative of the notebook entry. Therefore, there is an opportunity for tool designers to developer user tools for authoring artifacts of the research process, such as laboratory notebook pages, reports, and articles. Designs for such tools should consider the algorithmic needs of search, browse, and discovery algorithms, such as those described here. A round-trip design would supply the motive for authors to enter structured data when composing research artifacts. It would simultaneously create the mechanisms for automation support at additional steps of the research process, such as experimental design.

APPENDIX A: GRAPH VISUALIZATION TOOLS

The following graph visualization tools and references have been identified as readily accessible. These can be brought to bear to the analysis of the graphs described in the experiments. Although this is an incomplete list, it is representative of the current state-of-the-art for exploring and visualizing graphs computationally.

"Graph drawing: algorithms for the visualization of graphs"	(Di Battista 1999)
Adaptagrams	http://adaptagrams.sourceforge.net/
CiteSpace	http://cluster.cis.drexel.edu/~cchen/citespace/
Cytoscape	http://www.cytoscape.org/
Gephi	http://gephi.org/
GML	http://www.infosun.fim.uni-passau.de/Graphlet/GML/
Graphviz	http://www.graphviz.org/
GUESS	http://graphexploration.cond.org/
Igraph	http://igraph.sourceforge.net/
JUNG	http://jung.sourceforge.net/
LGL	http://bioinformatics.icmb.utexas.edu/lgl/
Maple	http://www.maplesoft.com/
MCL	http://micans.org/mcl/
NetDraw	http://www.analytictech.com/Netdraw/netdraw.htm
NodeXL	http://nodexl.codeplex.com/
OpenOrd	http://www.cs.sandia.gov/~smartin/software.html

Pajek	http://pajek.imfm.si/doku.php
Prefuse	http://prefuse.org/
SemanticNetSA	(Pellegrino and Chen 2008)
Siena	http://stat.gamma.rug.nl/siena.html
SoNIA	http://www.stanford.edu/group/sonia/
Topicscape	http://www.topicscape.com/
Tulip	http://www.tulip-software.org/
UCINET	http://www.analytictech.com/ucinet/
Visone	http://visone.info/

Table 6: Graph Visualization Tools

APPENDIX B: FIELD NOTES FROM DRUG DISCOVERY RESEARCHER INTERVIEW A

Transcript of Personal Notes

Section 1

- SAR pairwise analysis done all the time.
- Macroscopic overview.
- Full overview not available to people moving on and off a project.
    - Keep a few constant.
    - A few on and off.
- Time versus Number of Compounds
- Candidate found then backup candidate found.
- Key compounds from slides – slides stop at first candidate.
- Cutting the project – almost two different projects.
- Time and structural similarity – correlate the clusters with biological data.
    - "Does one cluster give you the desired biological properties more often than not?"
- Similarity not necessarily what he would consider similar.
- Big clusters of connectivity make sense – these are libraries.
- Expectations verified or not met during the exploration of the data.
- It can make you ask – "What did you get out of these clusters? – Should we have spent so much time there?"
- Some were related chemically but not structurally.

Section 2

Indegree

- How many compounds that we made like it?
- Left green 783 is the initial lead.
- 477 from 1st library.
- Observations match expected.
- Why this was not picked up?
- 317 – why did you not highlight – it is essentially the same.
- 007 – was key for nitrogen-biological reason.
- 261 = 606 + 477 first combination of
- What use?
    - Some things are kind of weird.
    - This in combination with the pairwise.

- Three fundamental changes were made in the project.
- Key compounds are found by looking at indegree, maybe depends on library chemistry.
- Indegree versus time – how much time in a cluster?

Outdegree

- Made a lot very similar to this one – to answer a very specific question.
- If outdegree is high and you are not asking a very specific compound?
    - Potential use in a design meeting.
    - Not currently doing.
- Maybe useful to use a map in a design meeting – Where are we? – Where do we need to be?
- Interesting dependent by chemists.
- Outdegree is really a pairwise analysis.

Betweenness

- 606 is pulled up by betweenness.
- Are the top betweenness interesting for any reason?
- 093 had a high out degree.
- **Add candidates to the data set of key compounds.**
- "I don't know why these are getting picked up."
- I would have thought those were connected by that compound. Maybe an artifact of the similarity score.
- Betweenness to identify singletons versus library compounds.
- Interesting because it

Section 3

- More useful for problematic projects.
- More useful for someone who had not been on the project before.
- Only going to work when the team recognized the key compounds.
- Highlighting interesting that have not been followed up on.

LIST OF REFERENCES

Adai, A. T., S. V. Date, et al. (2004). "LGL: Creating a Map of Protein Function with an Algorithm for Visualizing Very Large Biological Networks." Journal of Molecular Biology **340**(1): 179-190.

Allen, R. B. (2011). "Model-Oriented Scientific Research Reports." D-Lib Magazine **17**(5/6).

Arms, W. Y. and R. L. Larsen (2007). The Future of Scholarly Communication: Building the Infrastructure for Cyberscholarship, National Science Foundation and the Joint Information Systems Committee.

Arp, R. (2008). Scenario visualization : an evolutionary account of creative problem solving. Cambridge, Mass., MIT Press.

Atkins, D., S. Baker, et al. (2011). National Science Foundation Advisory Committee for Cyberinfrastructure Task Force on Data and Visualization Final Report, National Science Foundation.

Bao, Y., P. Bolotov, et al. (2008). "The Influenza Virus Resource at the National Center for Biotechnology Information." Journal of Virology **82**(2): 596-601.

Benger, W. (2009). "On Safari in the File Format Jungle--Why Can't You Visualize My Data?" Computing in Science and Engineering **11**(6): 98-102.

Bradley, J.-C., A. S. I. D. Lang, et al. (2011). Collaboration Using Open Notebook Science in Academia. Collaborative Computational Technologies for Biomedical Research. S. Ekins, M. A. Z. Hupcey and A. J. Williams, John Wiley & Sonds, Inc.**:** 425-452.

Bradley, J.-C., K. Mirza, et al. (2010). "Reaction Attempts: The UsefulChem Project." from http://dx.doi.org/10.1038/npre.2010.4416.1.

Brandes, U. and J. Lerner (2008). "Visual analysis of controversy in user-generated encyclopedias." Information Visualization **7**(1): 34-48.

Brunt, R. (2005). Some Aspects of Indexing in British Intelligence, 1939-1945. Covert and Overt: Recollecting and Connecting Intelligence Service and Information Science. Medford, NJ, Information Today, Inc.**:** 81 - 106.

Chen, C. (2006). "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature." Journal of the American Society for Information Science and Technology **57**(3): 359-377.

Chien, L., A. Tat, et al. (2008). Grand challenge award 2008: Support for diverse analytic techniques - nSpace2 and GeoTime visual analytics. Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on, Columbus, OH, IEEE.

Cohen, J. (2009). "Flu Researchers Train Sights On Novel Tricks of Novel H1N1." Science **324**(5929): 870-871.

Cohen, J. (2009). "Out of Mexico? Scientists Ponder Swine Flu's Origins." Science **324**(5928): 700-702.

de Nooy, W., A. Mrvar, et al. (2005). Exploratory Social Network Analysis with Pajek. New York, NY, Cambridge University Press.

Di Battista, G. (1999). Graph drawing: algorithms for the visualization of graphs, Prentice Hall.

Dylan, J. (2009, November 5). "Science Commons." from http://sciencecommons.org/about/science-commons-dylan-video/.

Fleming, L., S. Mingo, et al. (2007). "Collaborative Brokerage, Generative Creativity, and Creative Success." Administrative Science Quarterly **52**(3): 443-475.

Gantz, J. F., C. Chute, et al. (2008). The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011, IDC.

Garfield, E. (1972). "Citation Analysis as a Tool in Journal Evaluation: Journals can be ranked by frequency and impact of citations for science policy studies." Science **178**(4060): 449-526.

Garfield, E. (2009, November 18). "Eugene Garfield, Ph.D. Career Overview." from http://www.garfield.library.upenn.edu/overvu.html.

Garfield, E., A. I. Pudovkin, et al. (2003). "Why do we need algorithmic historiography?" Journal of the American Society for Information Science and Technology **54**(5): 400-412.

Görg, C., Z. Liu, et al. (2007). Jigsaw meets Blue Iguanodon - The VAST 2007 Contest. IEEE VAST '07, Sacramento, CA.

Grinstein, G., C. Plaisant, et al. (2008). VAST 2008 Challenge: Introducing mini-challenges. Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on, Columbus, OH, IEEE.

Hilderman, R. J. and H. J. Hamilton (2001). Knowledge Discovery and Measures of Interest, Kluwer Academic Publishers.

Keim, B. (2009). "Computer Program Self-Discovers Laws of Physics." Wired Science.

King, R. D., J. Rowland, et al. (2009). "The Automation of Science." Science **324**(5923): 85-89.

Kostoff, R. N. (2009). "A Systematic Approach to Alternative Medical Procedures." BioScience **59**(9): 734-735.

Kostoff, R. N., J. A. Block, et al. (2009). "Literature-related discovery." Annual Review of Information Science and Technology **43**(1): 1-71.

Larkey, S. V. (1949). "The Army Medical Library Research Project at the Welch Medical Library." Bulletin of the Medical Library Association **37**(2): 121-124.

Lin, X., H. D. White, et al. (2003). "Real-time author co-citation mapping for online searching." Information Processing \& Management **39**(5): 689-706.

MacEachren, A. M. (1995). How Maps Work: Representation, Visualization, and Design. New York, NY, The Guilford Press.

Martin, S., W. M. Brown, et al. (2011). OpenOrd: an open-source toolbox for large graph layout. Proceedings of VDA 2011 Conference on Visualization and Data Analysis 2011, San Francisco, CA.

Misawa, E., T. Russell, et al. (2009). Cyber-Enabled Discovery and Innovation (CDI). Arlington, VA, National Science Foundation.

Morris, C. M., M. Kimpton, et al. (2009). Fedora Commons and DSpace Foundation Join Together to Create DuraSpace™ Organization.

National Science Foundation Cyberinfrastructure Council (2007). Cyberinfrastructure Vision for 21st Century Discovery. Arlington, VA, National Science Foundation.

Page, L. (2001). Method for node ranking in a linked database. USA. **US 6,285,999 B1**.

Pan, C.-C., D. Pellegrino, et al. (2008). VAST 2008 Wiki Editors Mini Challenge - Identifying Social Networks using Wiki.viz. IEEE VAST '08, Columbus, OH.

Payne, J., J. Solomon, et al. (2008). Grand challenge award: Interactive visual analytics palantir: The future of analysis. Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on, Columbus, OH, IEEE.

Pellegrino, D., J.-C. Bradley, et al. (2011). Supporting scientific discovery through linkages of literature and data. Philadelphia, PA.

Pellegrino, D. and C. Chen (2008). Automatic hypothesis generation and evaluation by network structure content analysis and visualization. Annual U.S. Department of Homeland Security University Network Summit. Washington, DC.

Pellegrino, D. and C. Chen (2011). Data repository mapping for influenza protein sequence analysis. Proceedings of VDA 2011 Conference on Visualization and Data Analysis 2011, San Francisco, CA.

Pellegrino, D., C. Chen, et al. (2008). North-East Visualization and Analytics Center (NEVAC) Team Entry. VAST Challenge Portal, National Institute of Standards and Technology.

Pellegrino, D., C.-C. Pan, et al. (2008). Grand Challenge Award: Data Integration - Visualization and Collaboration in the VAST 2008 Challenge. Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on, Columbus, OH, IEEE.

Plaisant, C., J.-D. Fekete, et al. (2008). "Promoting Insight-Based Evaluation of Visualizations: From Contest to Benchmark Repository." IEEE Transactions on Visualization and Computer Graphics **14**(1): 120-134.

Plaisant, C., G. Grinstein, et al. (2008). "Evaluating Visual Analytics at the 2007 VAST Symposium Contest." IEEE Computer Graphics and Applications **28**(2): 12-21.

Schmidt, M. and H. Lipson (2009). "Distilling Free-Form Natural Laws from Experimental Data." Science **324**(5923): 81-85.

Schmidt, M. and H. Lipson. (2009, April 2). "Maybe Robots Dream of Electric Sheep, But Can They Do Science?", from http://nsfgov.http.internapcdn.net/nsfgov_vitalstream_com/podcast/lipson_schmidt.mp3.

SEASR. (2009, November 5). "About SEASR." from http://seasr.org/.

Smith, M. (2009). A Research Agenda for an Academic Research Library, Perspectives from MIT.

Søndergaard, T. F., J. Andersen, et al. (2003). "Documents and the communication of scientific and scholarly intformation: Revising and updating the UNISIST model." Journal of Documentation **59**(3): 278-320.

Stasko, J., C. Görg, et al. (2007). Jigsaw: Supporting Investigative Analysis through Interactive Visualization. IEEE VAST '07, Sacramento, CA.

Stasko, J., C. Gorg, et al. (2008). "Jigsaw: supporting investigative analysis through interactive visualization." Information Visualization **7**(2): 118-132.

Strickland, L. S. (2005). Knowledge Transfer: Information Science Shapes Intelligence in the Cold War Era. Covert and Overt: Recollecting and Connecting Intelligence Service and Information Science. R. V. Williams and B.-A. Lipetz. Medford, NJ, Information Today Inc.**:** 147-166.

Swanson, D. (2001). "ASIST Award of Merit Acceptance Speech: On the Fragmentation of Knowledge, the Connection Explosion, and Assembling Other People's Ideas." Bulletin of the American Society for Information Science and Technology **27**(3): 12-14.

Swanson, D. (2008). "Welcome to Arrowsmith 3.0." from http://d-swanson.uchicago.edu/.

Swanson, D. R. (1986). "FISH OIL, RAYNAUDS SYNDROME, AND UNDISCOVERED PUBLIC KNOWLEDGE." Perspectives in Biology and Medicine **30**(1): 7-18.

Swanson, D. R. (1986). "Undiscovered Public Knowledge." Library Quarterly **56**(2): 103-118.

Swanson, D. R. and N. R. Smalheiser (1999). "Implicit text linkages between Medline records: Using Arrowsmith as an aid to scientific discovery." Library Trends **48**(1): 48-59.

Thomas, J. J. and K. A. Cook (2005). Illuminating the Path: The Research and Development Agenda for Visual Analytics.

Waltz, D. and B. G. Buchanan (2009). "Automating Science." Science **324**(5923): 43-44.

Weaver, C. (2004). Building Highly-Coordinated Visualizations in Improvise, Austin, TX.

Whiting, M. A., W. Cowley, et al. (2006). Threat stream data generator: creating the known unknowns for test and evaluation of visual analytics tools. BELIV '06: Proceedings of the 2006 AVI workshop on BEyond time and errors, Venice, Italy, ACM.

Wikipedia contributors. (2009, November 5). "Open Notebook Science." from http://en.wikipedia.org/w/index.php?title=Open_Notebook_Science&oldid=321776139.

Williams, R. V. and B.-A. Lipetz, Eds. (2005). Covert and Overt: Recollecting and Connecting Intelligence Service and Information Science, Information Today, Inc.

Wilson, P. (1968). Two kinds of power: an essay on bibliographical control, Berkeley, University of California Press.

Zeldes, N. (2009). "Infoglut." IEEE Spectrum **46**(10): 30-55.